Statistics for Business

Course Stuff

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

This Draft: August 28, 2023.

Communication

- Lectures will take place in person at the Troias Building, Room **T106** (4 $\times$ 3 hours each)
- You can contact me either by e-mail (**pkonstantinou.aueb@gmail.com**) or by telephone (**+30 210 8203197**).
- I have a strong preference for e-mail (**pkonstantinou.aueb@gmail.com**) for the following reasons:
  1. I can respond whenever I find time to do so (I commit to do so withing two working days of the incoming message), whereas there is no guarantee that I am in my office every day of the week!!!
- All material (slides, assignments, etc.) related to the course are or will be posted at **https://eclass.aueb.gr/courses/MISC181/** which is OPEN to access (no registration is required)

Course Evaluation – I

- Course outline is available at: https://eclass.aueb.gr/modules/document/file.php/MISC181/Outline - Business Statistics 2020.pdf
- Main reading:
  - Newbold, P., Carlson, W.L. and Thorne, B. M. (2013) *Statistics for Business and Economics*, 8th edition, Essex: Pearson Education
  - Stock, J. and Watson, M. (2020) *Introduction to Econometrics*, 4th Global Edition, New York: Pearson (Ch. 1 – Ch.4)
- **Course Assessment**:

Course Evaluation – II

- Weekly Assignments (30%) $\mapsto$ **pkonstantinou.aueb@gmail.com**. Anything sent to ***pkonstantinou@aueb.gr*** (my institutional e-mail address) will be *lost*. The answers to the assignments will have to be either typed or scanned (but always pdf files). **DO NOT SEND PICTURES** – they are too large and might not get through.
- Written Examination (70%) – dates will be announced.

Statistics for Business

Background: Descriptive Statistics

Panagiotis Th. Konstantinou

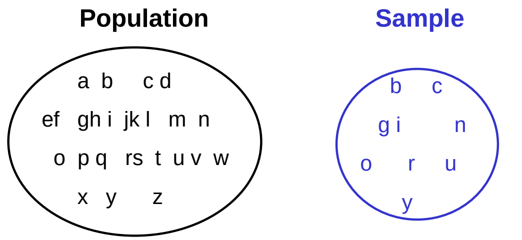
MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: July 15, 2015. **This Draft**: August 30, 2023.

Key Concepts

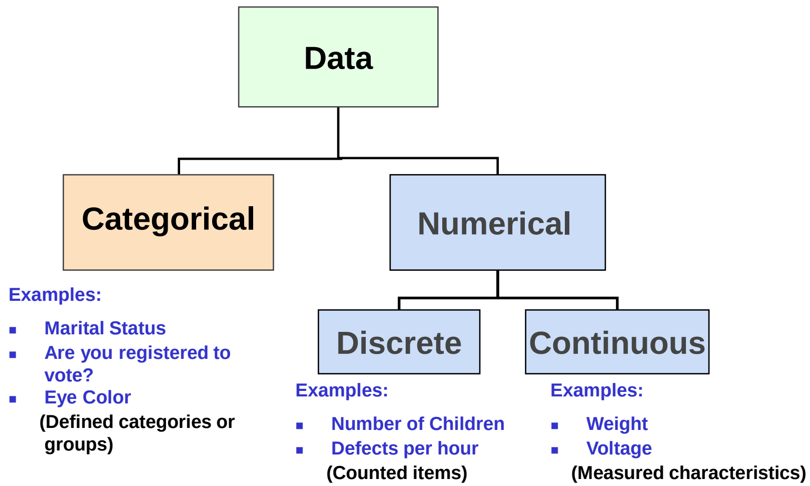
- A **population** is the collection of all items of interest or under investigation (N represents the population size)
- A **sample** is an observed subset of the population (n represents the sample size)
- A **parameter** is a specific characteristic of a population
- A **statistic** is a specific characteristic of a sample



Values calculated using population data are called parameters

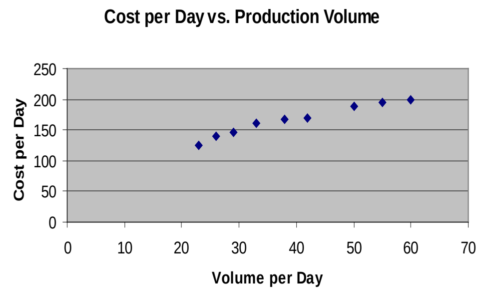
Values computed from sample data are called statistics

Data Types



Relationships Between Variables

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46.5	55	27.5	**129**
Bonds	32.0	44	19.0	**95**
CD	15.5	20	13.5	**49**
Savings	16.0	28	7.0	**51**
Total	**110.0**	**147**	**67.0**	**324**

Describing Data Numerically

Describing Data Numerically

- Central Tendency
  - Arithmetic Mean
  - Median
  - Mode
- Variation
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Measures of Central Tendency

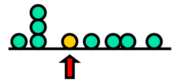
Central Tendency

- Mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

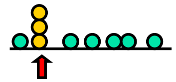
Arithmetic average

- Median



Midpoint of ranked values

- Mode



Most frequently observed value

- Median position $\frac{n+1}{2}$ position in the ordered data
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

Measures of Central Tendency

Example

House Prices	
	\$2,000,000
	500,000
	300,000
	100,000
	100,000
Sum	\$3,000,000

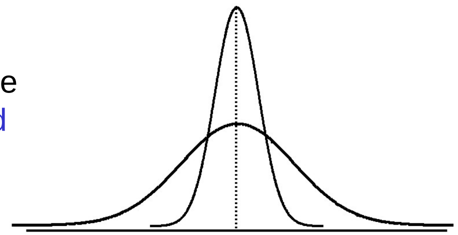
- **Mean**: \$3,000,000/5 = \$600,000
- **Median**: middle value of ranked data = \$300,000
- **Mode**: most frequent value = \$100,000

Measures of Variability

Variation

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

- Measures of variation give information on the spread or variability of the data values.



Same center, different variation

Variance

- ***Population Variance*:** Average of squared deviations of values from the mean

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

where
  - $\mu =$ population mean
  - $N =$ population size
  - $X_i = i-$th value of the variable X

- ***Sample Variance*:** Average (approximately) of squared deviations of values from the sample mean:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where
  - $\bar{x} =$ sample mean/average
  - $n =$ sample size
  - $x_i = i-$th value of the variable X

Standard Deviation

- ***Population Standard Deviation*:** Most commonly used measure of variation
  - Shows variation about the mean
  - Has the *same units as the original data*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

- ***Sample Standard Deviation*:** Most commonly used measure of variation
  - Shows variation about the *sample* mean
  - Has the *same units as the original data*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard Deviation

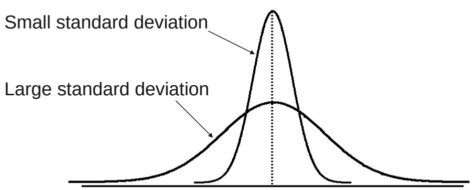
Example: Sample Standard Deviation Computation

- Sample Data (x_i) : 10 12 14 15 17 18 18 24
- $n = 8$ and sample mean $= \bar{x} = 16$
- So the standard deviation is

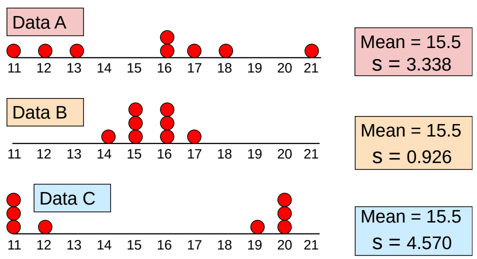
$$\begin{aligned} s &= \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}} \\ &= \sqrt{\frac{126}{7}} = 4.2426 \end{aligned}$$

- This is a measure of the "**average**" scatter around the (sample) mean.

Comparing Standard Deviations

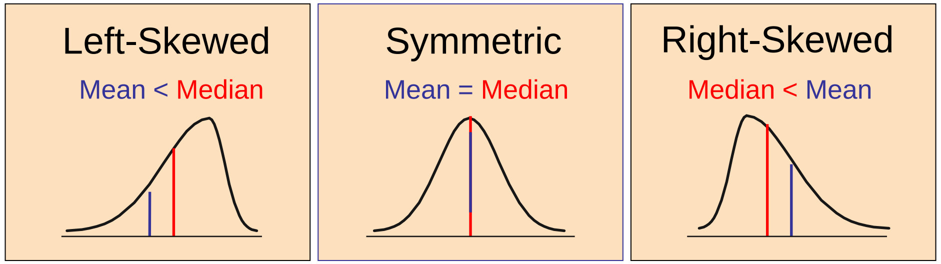


- The smaller the standard deviation, the more concentrated are the values around the mean.



- Same mean, different standard deviations.

Shape of a Distribution



- Describes how data are distributed
- Measures of **shape**:
  - Symmetric or skewed
  - Left = Negative (mass of distr. concentrated on the right of figure); Right = Positive (mass of distr. concentrated on the left of figure).

$$SK = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Coefficient of Variation

- Measures relative variation and is always in percentage (%)
- Shows variation **relative to mean**
- Can be used to compare two or more sets of data **measured in different units**

$$CV = \left(\frac{s_x}{\bar{x}}\right) \cdot 100\%$$

- **Stock A**:
  - Avg price last year = \$50
  - Standard deviation = \$5

$$CV_A = \left(\frac{\$5}{\$50}\right) \cdot 100\% = 10\%$$

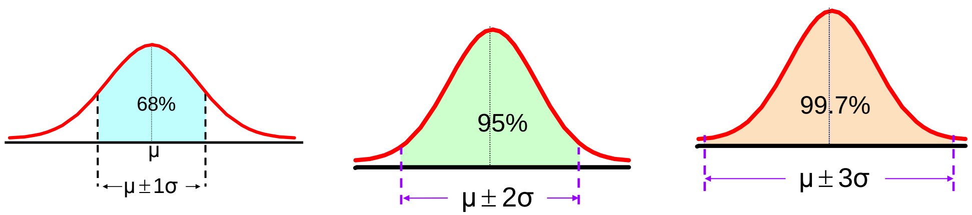
- **Stock B**:
  - Avg. price last year = \$100
  - Standard deviation = \$5

$$CV_B = \left(\frac{\$5}{\$100}\right) \cdot 100\% = 5\%$$

- Both stocks have the same standard deviation, but stock B is less variable relative to its price

The Empirical Rule

If the data distribution is bell-shaped, then the interval:



- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample
- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains almost all (about 99.7%) of the values in the population or the sample.

Covariance

- The covariance measures the strength of the linear relationship between **two variables**
- The ***population covariance***:

$$\text{Cov}(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}.$$

- The ***sample covariance***:

$$\widehat{\text{Cov}(x, y)} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

- Only concerned with the strength of the relationship
- No causal effect is implied
  - $\text{Cov}(x, y) > 0$, x and y tend to move in the ***same*** direction
  - $\text{Cov}(x, y) < 0$, x and y tend to move in ***opposite*** directions

Correlation Coefficients

- The correlation coefficient measures the relative strength of the linear relationship between **two variables**
- The ***population correlation coefficient***:

$$\text{Corr}(X, Y) = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

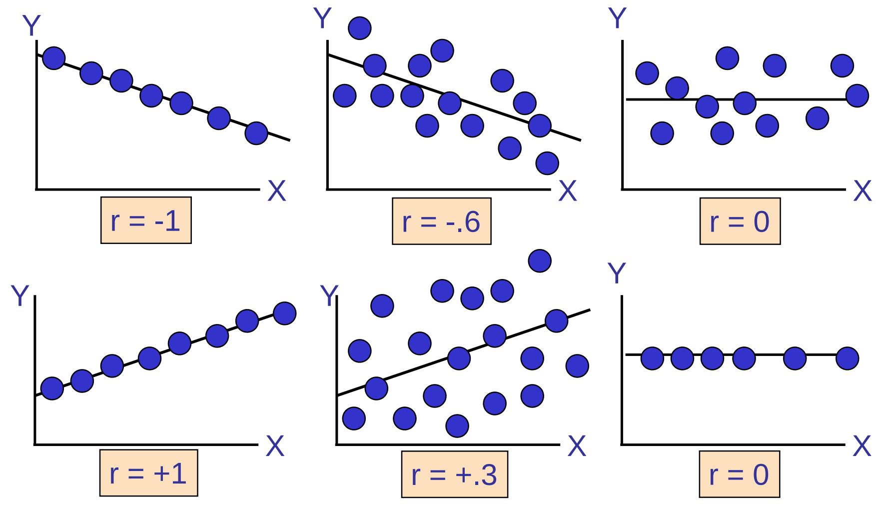
- The ***sample correlation coefficient***:

$$\widehat{\text{Corr}(x, y)} = r_{xy} = \frac{\widehat{\text{Cov}(x, y)}}{s_x s_y}.$$

- Unit free and ranges between -1 and 1
  - The closer to -1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker any positive linear relationship

Correlation Coefficients

Examples



Statistics for Business

Elements of Probability Theory

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

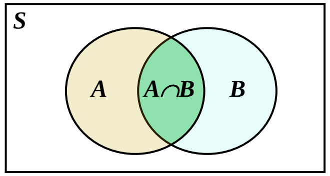
Athens University of Economics and Business

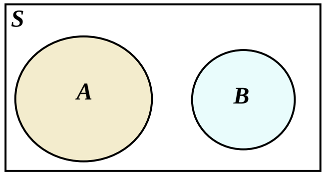
First Draft: July 15, 2015. **This Draft**: August 28, 2023.

Important Terms in Probability – I

- ***Random Experiment*** – it is a process leading to an uncertain outcome
- ***Basic Outcome*** (S_i) – a possible outcome (the most basic one) of a random experiment
- ***Sample Space*** (S) – the collection of all possible (basic) outcomes of a random experiment
- ***Event*** A – is any subset of basic outcomes from the sample space ($A \subseteq S$). This is our object of interest here – among other things.

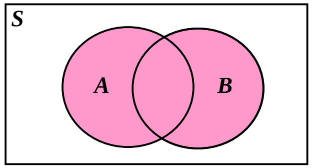
Important Terms in Probability – II

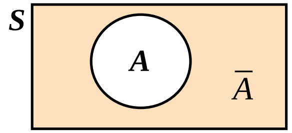




- ***Intersection of Events*** – If A and B are two events in a sample space S, then their intersection, $A \cap B$, is the set of all outcomes in S that belong to **both** A and B
- We say that A and B are ***Mutually Exclusive Events*** if they have no basic outcomes in common i.e., the set $A \cap B$ is empty ($\varnothing$)

Important Terms in Probability – III





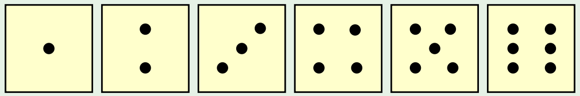
- ***Union of Events*** – If A and B are two events in a sample space S, then their union, $A \cup B$, is the set of all outcomes in S that belong to either A or B
- The ***Complement*** of an event A is the set of all basic outcomes in the sample space that do not belong to A. The complement is denoted $\bar{A}$ or A^c.

Important Terms in Probability – IV

- Events $E_1, E_2, \ldots, E_k$ are ***Collectively Exhaustive*** events if $E_1 \cup E_2 \cup \ldots \cup E_k = S$, i.e., the events completely cover the sample space.

Examples

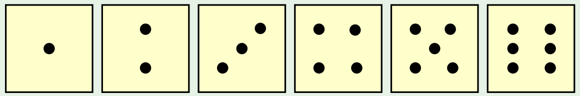
Let the ***Sample Space*** be the collection of all possible outcomes of rolling one die $S = \{1, 2, 3, 4, 5, 6\}$.



- Let A be the event "Number rolled is even": $A = \{2, 4, 6\}$
- Let B be the event "Number rolled is at least 4" : $B = \{4, 5, 6\}$
- ***Mutually exclusive***: A and B are **not** mutually exclusive. The outcomes 4 and 6 are common to both.

Important Terms in Probability – V

Examples (Continued)



$A = \{2, 4, 6\}$ $\quad$ $B = \{4, 5, 6\}$

- ***Collectively exhaustive***: A and B are **not** collectively exhaustive. $A \cup B$ does not contain 1 or 3.
- ***Complements***: $\bar{A} = \{1, 3, 5\}$ and $\bar{B} = \{1, 2, 3\}$
- ***Intersections***: $A \cap B = \{4, 6\}$; $\bar{A} \cap B = \{5\}$; $A \cap \bar{B} = \{2\}$; $\bar{A} \cap \bar{B} = \{1, 3\}$.
- ***Unions***: $A \cup B = \{2, 4, 5, 6\}$; $A \cup \bar{A} = \{1, 2, 3, 4, 5, 6\} = S$.

Assessing Probability – I

- ***Probability*** – the chance that an uncertain event A will occur is always between 0 and 1.

$$\underbrace{0}_{\text{Impossible}} \leq \Pr(A) \leq \underbrace{1}_{\text{Certain}}$$

- There are three approaches to assessing the probability of an uncertain event:

Assessing Probability – II

1. ***Classical Definition of Probability***:

$$\text{Probability of an event } A = \frac{N_A}{N}$$
$$= \frac{\text{number of outcomes that satisfy the event } A}{\text{total number of outcomes in the sample space } S}$$

   - Assumes all outcomes in the sample space are equally likely to occur.
   - ***Example***: Consider the experiment of tossing 2 coins. The sample space is $S = \{HH, HT, TH, TT\}$.
   - Event $A = \{\text{one } T\} = \{TH, HT\}$. Hence $\Pr(A) = 0.5$ – assuming that all basic outcomes are equally likely.
   - Event $B = \{\text{at least one } T\} = \{TH, HT, TT\}$. So $\Pr(B) = 0.75$.

Assessing Probability – III

2. ***Probability as Relative Frequency***:

$$\text{Probability of an event } A = \frac{n_A}{n}$$
$$= \frac{\text{number of events in the population that satisfy event } A}{\text{total number of events in the population}}$$

   - The limit of the proportion of times that an event A occurs in a large number of trials, n.

Assessing Probability – IV

3. ***Subjective Probability***: an individual has opinion or belief about the probability of occurrence of A.
   - When economic conditions or a company's circumstances change rapidly, it might be inappropriate to assign probabilities based solely on historical data
   - We can use any data available as well as our experience and intuition, but ultimately a probability value should express our degree of belief that the experimental outcome will occur.

Measuring Outcomes – I

Classical Definition of Probability

- ***Basic Rule of Counting***: If an experiment consists of a sequence of k steps in which there are n_1 possible results for the first step, n_2 possible results for the second step, and so on, then the total number of experimental outcomes is given by $(n_1)(n_2)\ldots(n_k)$ – tree diagram...

Measuring Outcomes – II

Classical Definition of Probability

- ***Counting Rule for Combinations*** (Number of Combinations of n Objects taken k at a time): A second useful counting rule enables us to count the number of experimental outcomes when k objects are to be selected from a set of n objects (the ordering does not matter)

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!},$$

where $n! = n(n-1)(n-2)\ldots(2)(1)$ and $0! = 1$.

Measuring Outcomes – III

Classical Definition of Probability

- ***Example***: Suppose we flip three coins. How many are the possible combinations with (exactly) 1 T?

$$C_1^3 = \binom{3}{1} = \frac{3!}{1!(3-1)!} = 3.$$

- ***Example***: Suppose we flip three coins. How many are the possible combinations with *at least* 1T?
- ***Example***: Suppose that there are two groups of questions. Group A with 6 questions and group B with 4 questions. How many are the possible half-a-dozens we can put together?

$$n = 6 + 4 = 10; \; C_6^{10} = \binom{10}{6} = \frac{10!}{6!(10-6)!} = 210.$$

Measuring Outcomes – IV

Classical Definition of Probability

- ***Example***: How many possible half-a-dozens we can put together, preserving the ratio 4 : 2?

$$\binom{6}{4} \times \binom{4}{2} = 15 \times 6 = 90.$$

- **Probability**: What is the probability of selecting a particular half-a-dozen (with ratio 4 : 2), when we choose at random? Using the classical definition of probability

$$\frac{90}{210} = 0.4286$$

Measuring Outcomes – V

Classical Definition of Probability

- ***Counting Rule for Permutations*** (Number of Permutations of n Objects taken k at a time): A third useful counting rule enables us to count the number of experimental outcomes when k objects are to be selected from a set of n objects, **where the order of selection is important**

$$P_k^n = \frac{n!}{(n-k)!}.$$

Measuring Outcomes – VI

Classical Definition of Probability

- ***Example***: How many 3-digit lock combinations can we make from the numbers 1, 2, 3, and 4?
The order of the choice is important! So

$$P_3^4 = \frac{4!}{1!} = 4! = 4(3)(2)(1) = 24.$$

- ***Example***: Let the characters A, B, Γ. In how many ways can we combine them in making triads?

$$P_3^3 = \frac{3!}{0!} = 3! = 3(2)(1) = 6.$$

These are: $AB\Gamma, A\Gamma B, BA\Gamma, B\Gamma A, \Gamma AB$,and ΓBA.

Measuring Outcomes – VII

Classical Definition of Probability

- ***Example***: Let the characters A, B, Γ, Δ, E. In how many ways is it possible to combine them into pairs?

* If the order matters, we may have

$$P_2^5 = \frac{5!}{3!} = (5)(4) = 20.$$

* If the order does not matters, we may choose pairs

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2!3!} = 10$$

Probability Axioms

- The following ***Axioms*** hold

1. If A is any event in the sample space S, then

$$0 \leq \Pr(A) \leq 1.$$

2. Let A be an event in S, and let S_i denote the basic outcomes. Then

$$\Pr(A) = \sum_{\text{all } S_i \text{ in } A} \Pr(S_i).$$

3. $\Pr(S) = 1.$

Probability Rules – I

- The ***Complement Rule***:

$$\Pr(\bar{A}) = 1 - \Pr(A) \text{ [i.e., } \Pr(A) + \Pr(\bar{A}) = 1].$$

- The ***Addition Rule***: The probability of the union of two events is

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

- Probabilities and joint probabilities for two events A and B are summarized in the following table:

	B	$\bar{B}$	
A	$\Pr(A \cap B)$	$\Pr(A \cap \bar{B})$	$\Pr(A)$
$\bar{A}$	$\Pr(\bar{A} \cap B)$	$\Pr(\bar{A} \cap \bar{B})$	$\Pr(\bar{A})$
	$\Pr(B)$	$\Pr(\bar{B})$	$\Pr(S) = 1$

Probability Rules – II

Example (Addition Rule)

Consider a standard deck of 52 cards, with four suits ♡♣♢♠. Let event A = card is an Ace and event B = card is from a red suit.

Pr(Red ∪ Ace) = Pr(Red) + Pr(Ace) - Pr(Red ∩ Ace)

= 26/52 + 4/52 - 2/52 = 28/52

Don't count the two red aces twice!

Type	**Color**		**Total**
	Red	**Black**	
Ace	**2**	**2**	**4**
Non-Ace	**24**	**24**	**48**
Total	**26**	**26**	**52**

Conditional Probability – I

- A ***conditional probability*** is the probability of one event, given that another event has occurred:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \text{ (if } \Pr(B) > 0);$$

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \text{ (if } \Pr(A) > 0)$$

Conditional Probability – II

Example (Conditional Probability)

Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC? [$\Pr(CD|AC) = ?$]

	CD	**No CD**	**Total**
AC	**.2**	**.5**	**.7**
No AC	**.2**	**.1**	**.3**
Total	**.4**	**.6**	**1.0**

$$\Pr(\text{CD}|\text{AC}) = \frac{\Pr(\text{CD} \cap \text{AC})}{\Pr(\text{AC})} = \frac{.2}{.7} = .2857$$

Multiplication Rule

- The ***Multiplication Rule*** for two events A and B:

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$

Example (Multiplication Rule)

Pr(Red ∩ Ace) = Pr(Red| Ace)Pr(Ace)

$$= \left(\frac{2}{4}\right)\left(\frac{4}{52}\right) = \frac{2}{52}$$

$$= \frac{\text{number of cards that are red and ace}}{\text{total number of cards}} = \frac{2}{52}$$

Type	Color: Red	Color: Black	Total
Ace	2	2	4
Non-Ace	24	24	48
Total	26	26	52

Statistical Independence – I

- Two events are ***statistically independent*** if and only if:

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

  - Events A and B are independent when the probability of one event is not affected by the other event.
  - If A and B are independent, then

$$\begin{aligned} \Pr(A|B) &= \Pr(A), \text{ if } \Pr(B) > 0; \\ \Pr(B|A) &= \Pr(B), \text{ if } \Pr(A) > 0. \end{aligned}$$

Statistical Independence – II

Example (Statistical Independence)

Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both. Are the events AC and CD statistically independent?

	CD	No CD	Total
AC	.2	.5	.7
No AC	.2	.1	.3
Total	.4	.6	1.0

P(AC ∩ CD) = 0.2

$$\left.\begin{array}{l}\mathbf{P(AC) = 0.7} \\ \mathbf{P(CD) = 0.4}\end{array}\right\} \mathbf{P(AC)P(CD) = (0.7)(0.4) = 0.28}$$

P(AC ∩ CD) = 0.2 ≠ P(AC)P(CD) = 0.28
So the two events are not statistically independent

Statistical Independence – III

Remark (Exclussive Events and Statistical Independence)

*Let two events A and B with $\Pr(A) > 0$ and $\Pr(B) > 0$ which are mutually exclusive. Are A and B independent? **NO**!*

To see this use a Venn diagram and the formula of conditional probability (or the multiplication rule).

- If one mutually exclusive event is known to occur, the other cannot occur; thus, the probability of the other event occurring is reduced to zero (and they are therefore dependent).

Examples – I

- **Example 1.** In a certain population, 10% of the people can be classified as being high risk for a heart attack. Three people are randomly selected from this population. What is the probability that exactly one of the three are high risk?

▶ Define H: high risk, and N: not high risk. Then

$$\Pr(\text{exactly one high risk}) = \Pr(HNN) + \Pr(NHN) + \Pr(NNH) =$$
$$= \Pr(H)\Pr(N)\Pr(N) + \Pr(N)\Pr(H)\Pr(N) + \Pr(N)\Pr(N)\Pr(H)$$
$$= (.1)(.9)(.9) + (.9)(.1)(.9) + (.9)(.9)(.1) = 3(.1)(.9)^2 = .243$$

Examples – II

- **Example 2.** Suppose we have additional information in the previous example. We know that only 49% of the population are female. Also, of the female patients, 8% are high risk. A single person is selected at random. What is the probability that it is a high risk female?

▶ Define H: high risk, and F: female. From the example, $\Pr(F) = .49$ and $\Pr(H|F) = .08$. Using the Multiplication Rule:

$$\Pr(\text{high risk female}) = \Pr(H \cap F)$$
$$= \Pr(F)\Pr(H|F) = .49(.08) = .0392$$

Statistics for Business

Random Variables and Probability Distributions, Special Discrete and Continuous Probability Distributions

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: July 15, 2045. **This Draft**: August 28, 2023.

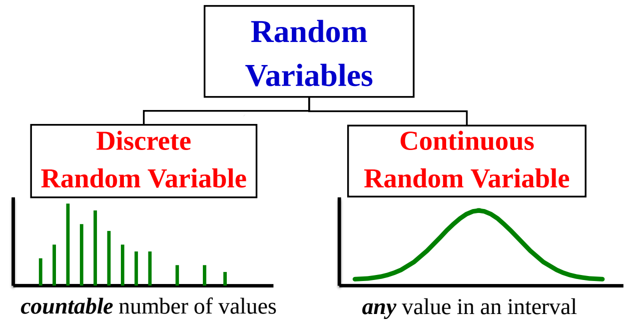
Random Variables – I

Basics

Definition

A ***random variable*** X is a a function or rule that assigns a **number** to each outcome of an experiment.

Think of this as the numerical summary of a random outcome.



Random Variables – II

Basics

Examples

- X = GPA for a randomly selected student
- X = number of contracts a shipping company has pending at a randomly selected month of the year
- X = number on the upper face of a randomly tossed die
- X = the price of crude oil during a randomly selected month.

Discrete Random Variables

- A ***discrete random variable*** can only take on a countable number of values

Examples

- Roll a die twice. Let X be the number of times 4 comes up:
  - then X could be 0, 1, or 2 times
- Toss a coin 5 times. Let X be the number of heads:
  - then $X = 0, 1, 2, 3, 4$, or 5

Discrete Probability Distributions – I

- The ***probability distribution*** for a ***discrete random variable*** X resembles the relative frequency distributions. It is a graph, table or formula that gives the possible values of X and the probability $P(X = x)$ associated with each value.
- This must satisfy
  1. $0 \leq P(x) \leq 1$, for all x.
  2. $\sum_{\text{all } x} P(x) = 1$, the individual probabilities sum to 1.
- The ***cumulative probability function***, denoted by $F(x_0)$, shows the probability that X is less than or equal to a particular value, x_0 :

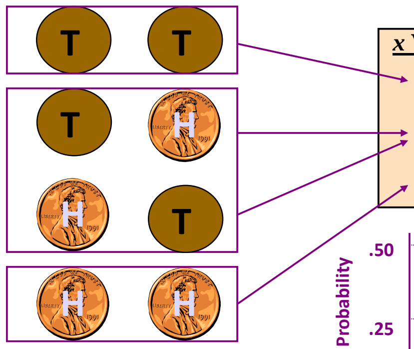
$$F(x_0) = \Pr(X \leq x_0) = \sum_{x \leq x_0} P(x)$$

Discrete Probability Distributions – II

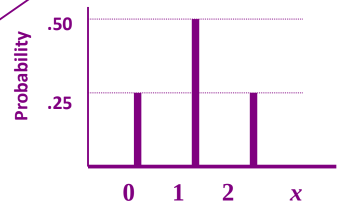
- **Random Experiment**: Toss 2 Coins. Let (the random variable) $X = \#$ heads.

Discrete Probability Distributions – III

4 possible outcomes **Probability Distribution**



x Value	Probability	Cum. Prob.
0	1/4 = .25	1/4 = .25
1	2/4 = .50	3/4 = .75
2	1/4 = .25	4/4 = 1.00



- **Random Experiment**: Let the random variable S be the number of days it will snow in the last week of January

Discrete Probability Distributions – IV

(cumulative) Probability distribution of S								
Outcome	0	1	2	3	4	5	6	7
Probability	0.20	0.25	0.20	0.15	0.10	0.05	0.04	0.01
CDF	0.20	0.45	0.65	0.80	0.90	0.95	0.99	1.00

Moments of Discrete Prob. Distributions – I

- ***Expected Value*** (or ***mean***) of a discrete distribution (*weighted average*)

$$\mu_X = \mathrm{E}(X) = \sum_{\text{all } x} x \cdot P(x).$$

- ***Variance*** of a discrete random variable X (*weighted average...*)

$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}\left[(X-\mu_X)^2\right] = \sum_{\text{all } x} (x-\mu_X)^2 \cdot P(x)$$

- ***Standard Deviation*** of a discrete random variable X

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{\text{all } x} (x-\mu)^2 P(x)}$$

Moments of Discrete Prob. Distributions – II

Example

Consider the experiment of tossing 2 coins, and $X = \#$ of heads. Then

$$\begin{aligned} \mu &= \mathrm{E}(X) = \sum_x xP(x) \\ &= (0 \times 0.25) + (1 \times 0.50) + (2 \times 0.25) = 1 \end{aligned}$$

$$\begin{aligned} \sigma &= \sqrt{\sum_x (x-\mu)^2 P(x)} \\ &= \sqrt{(0-1)^2(.25) + (1-1)^2(.50) + (2-1)^2(.25)} \\ &= \sqrt{.50} = 0.707 \end{aligned}$$

Moments of Discrete Prob. Distributions – III

Example (Number of days it will snow in January)

$\mu_S = \mathrm{E}(S) = \sum_s s \cdot P(s) =$
$= 0{\cdot}0.2{+}1{\cdot}0.25{+}2{\cdot}0.2{+}3{\cdot}0.15{+}4{\cdot}0.1{+}5{\cdot}0.05{+}6{\cdot}0.04{+}7{\cdot}0.01 = 2.06$

$\sigma_S^2 = \mathrm{Var}(S) = \sum_s (s - \mathrm{E}(S))^2 \cdot P(s) =$
$= (0{-}2.06)^2{\cdot}0.2{+}(1{-}2.06)^2{\cdot}0.25{+}(2{-}2.06)^2{\cdot}0.2{+}(3{-}2.06)^2{\cdot}0.15$
$+(4-2.06)^2 \cdot 0.1 + (5-2.06)^2 \cdot 0.05 + (6-2.06)^2 \cdot 0.04$
$+(7-2.06)^2 \cdot 0.01 = 2.94$

Remark (Rules for Moments)

Let a and b be any constants and let $Y = a + bX$. Then

$$\begin{aligned} \mathrm{E}[a+bX] &= a + b\mathrm{E}[X] = a + b\mu_x \\ \mathrm{Var}[a+bX] &= b^2\mathrm{Var}[X] = b^2\sigma_x^2 \Rightarrow \sigma_Y = |b|\sigma_x \end{aligned}$$

- The above imply that $\mathrm{E}[a] = a$ and $\mathrm{Var}[a] = 0$

Prob. Density and Distribution Function – I

- The ***probability density function*** (or *pdf*), $f(x)$, of continuous random variable X has the following properties

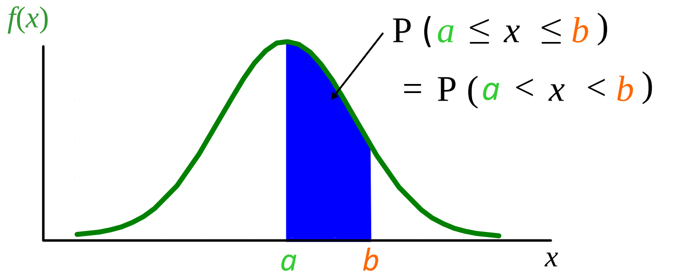
1. $f(x) > 0$ for all values of x (x takes a range of values, $\mathbb{R}_X$).
2. The area under the probability density function $f(x)$ over all values of the random variable X is equal to 1 (recall that $\sum_{\text{all } x} P(x) = 1$ for discrete r.v.)

$$\int_{\mathbb{R}_X} f(x)dx = 1.$$

Prob. Density and Distribution Function – II

3. The probability that X lies between two values is the area under the density function graph between the two values:

$$\Pr(a \le X \le b) = \Pr(a < X < b) = \int_a^b f(x)dx$$



Note that the probability of any individual value is zero

Prob. Density and Distribution Function – III

4. The ***cumulative density function*** (or ***distribution function***) $F(x_0)$, which expresses the probability that X does not exceed the value of x_0, is the area under the probability density function $f(x)$ from the minimum x value up to x_0

$$F(x_0) = \int_{x_{\min}}^{x_0} f(x)dx.$$

5. It follows that

$$\Pr(a \le X \le b) = \Pr(a < X < b) = F(b) - F(a)$$

Moments of Continuous Distributions – I

- ***Expected Value*** (or ***mean***) of a continuous distribution

$$\mu_X = \mathrm{E}(X) = \int_{\mathbb{R}_X} xf(x)dx.$$

- ***Variance*** of a continuous random variable X

$$\sigma_X^2 = \mathrm{Var}(X) = \int_{\mathbb{R}_X} (x-\mu_X)^2 f(x)dx$$

- ***Standard Deviation*** of a continuous random variable X

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\int_{\mathbb{R}_X} (x-\mu_X)^2 f(x)dx}$$

Moments of Continuous Distributions – II

Remark (Rules for Moments Apply)

Let c and d be any constants and let $Y = c + dX$. Then

$$\begin{aligned} \mathrm{E}[c+dX] &= c + d\mathrm{E}[X] = c + d\mu_x \\ \mathrm{Var}[c+dX] &= d^2\mathrm{Var}[X] = d^2\sigma_x^2 \Rightarrow \sigma_Y = |d|\sigma_x \end{aligned}$$

Remark (Standardized Random Variable)

*An **important special case** of the previous results is*

$$Z = \frac{X-\mu_x}{\sigma_x},$$

$$\textit{for which} \quad : \quad \begin{matrix} \mathrm{E}(Z) = 0 \\ \mathrm{Var}(Z) = 1 \end{matrix}.$$

Bernoulli Distribution

- Consider only two outcomes: "*success*" or "*failure*". Let p denote the probability of success, and $1-p$ be the probability of failure.
- Define random variable X: $x = 1$ if success, $x = 0$ if failure.
- Then the *Bernoulli probability function* is

$$P(X=0) = (1-p) \text{ and } P(X=1) = p$$

- Moreover:

$$\begin{aligned} \mu_X &= \mathrm{E}(X) = \sum_{\text{all } x} x \cdot P(x) = 0\cdot(1-p) + 1\cdot p = p \\ \sigma_X^2 &= \mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2] = \sum_{\text{all } x} (x-\mu_X)^2 \cdot P(x) \\ &= (0-p)^2(1-p) + (1-p)^2 p = p(1-p) \end{aligned}$$

Binomial Distribution – I

- A fixed number of observations, n
  - e.g., 15 tosses of a coin; ten light bulbs taken from a warehouse
- Two mutually exclusive and collectively exhaustive categories
  - e.g., head or tail in each toss of a coin; defective or not defective light bulb
  - Generally called "*success*" and "*failure*"
  - Probability of success is p, probability of failure is $1-p$
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin
- Observations are independent
  - The outcome of one observation does not affect the outcome of the other

Binomial Distribution – II

- Examples:
  - A manufacturing plant labels items as either defective or acceptable
  - A firm bidding for contracts will either get a contract or not
  - A marketing research firm receives survey responses of “yes I will buy” or “no I will not”
  - New job applicants either accept the offer or reject it
- To calculate the probability associated with each value we use combinatorics:

$$P(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}; \quad x = 0, 1, 2, \ldots, n$$

Binomial Distribution – III

- $P(x)$ = probability of x successes in n trials, with probability of success p on each trial; x = number of ‘successes’ in sample (nr. of trials n); $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$

Example

What is the probability of one success in five observations if the probability of success is 0.1?

- Here $x = 1, n = 5$, and $p = 0.1$. So

$$\begin{aligned} P(x &= 1) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\ &= \frac{5!}{1!(5-1)!}(0.1)^1(1-0.1)^{5-1} = 5(0.1)(0.9)^4 = 0.32805 \end{aligned}$$

Binomial Distribution

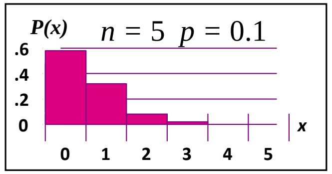
Moments and Shape

$$\begin{aligned} \mu &= \mathrm{E}(X) = np \\ \sigma^2 &= \mathrm{Var}(X) = np(1-p) \Rightarrow \sigma = \sqrt{np(1-p)} \end{aligned}$$

- The shape of the binomial distr. depends on the values of p and n

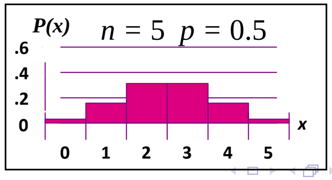
$\mu = np = (5)(0.1) = 0.5$

$\sigma = \sqrt{np(1-p)} = \sqrt{(5)(0.1)(1-0.1)} = 0.6708$



$\mu = np = (5)(0.5) = 2.5$

$\sigma = \sqrt{np(1-p)} = \sqrt{(5)(0.5)(1-0.5)} = 1.118$



Normal Distribution – I

- The *normal distribution* is the most important of all probability distributions. The probability density function of a **normal random variable** is given by

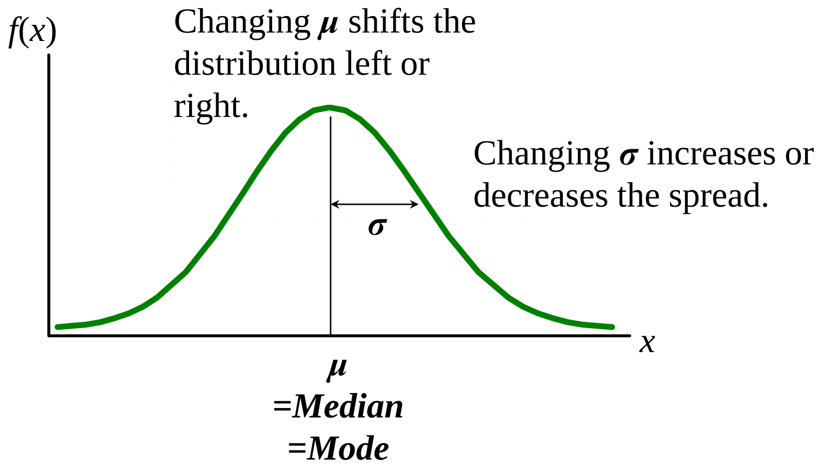
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad -\infty < x < +\infty,$$

and we usually write $X \sim N(\mu_x, \sigma_x^2)$

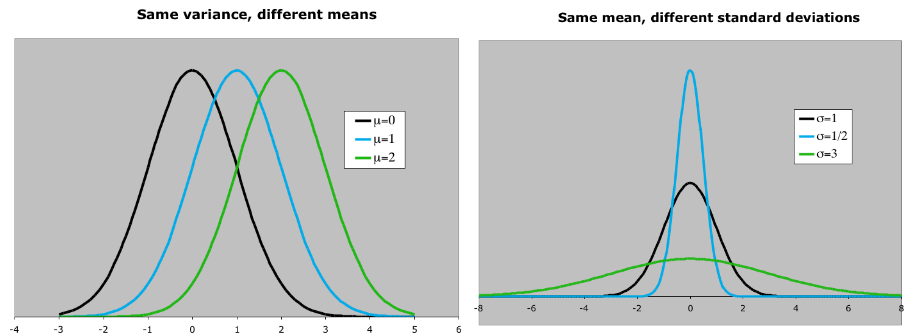
  - The normal distribution closely approximates the probability distributions of a wide range of random variables
  - Distributions of sample means approach a normal distribution given a “large” sample size
  - Computations of probabilities are direct and elegant

Normal Distribution – II

- The shape and location of the normal curve changes as the mean (μ) and standard deviation (σ) change



Normal Distribution – III



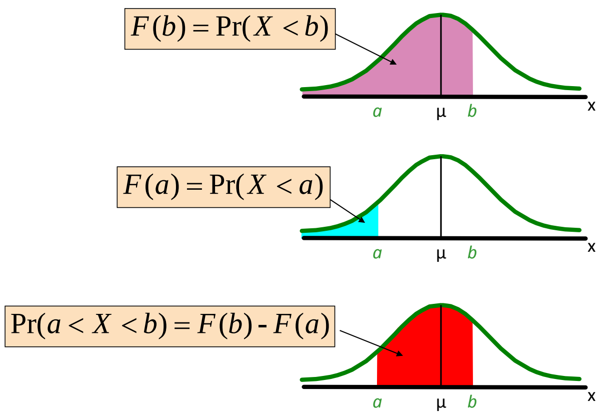
- For a normal random variable X with mean μ and variance σ^2, i.e., $X \sim N(\mu, \sigma^2)$), the cumulative distribution function is

$$F(x_0) = \Pr(X \leq x_0),$$

Normal Distribution – IV

while the probability for a range of values is measured by the area under the curve

$$\Pr(a < X < b) = F(b) - F(a)$$



Normal Distribution – V

- Any normal distribution (with any mean and variance combination) can be transformed into the standardized normal distribution (Z), with mean 0 and variance 1:

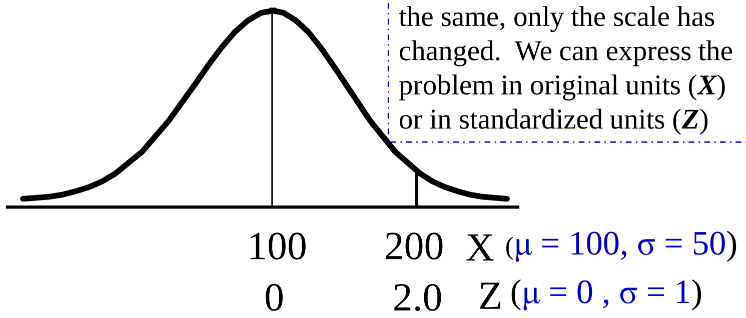
$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- **Example**: If $X \sim N(100, 50^2)$, the Z value for $X = 200$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2$$

This says that $X = 200$ is two standard deviations (2 increments of 50 units) above the mean of 100.

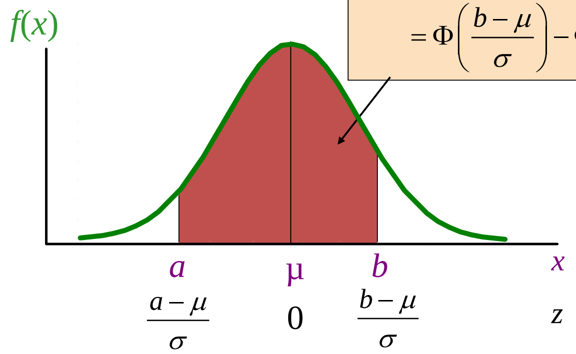
Normal Distribution – VI



Note that the distribution is the same, only the scale has changed. We can express the problem in original units (***X***) or in standardized units (***Z***)

Finding Normal Probabilities – I

$$\Pr(a < X < b) = \Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

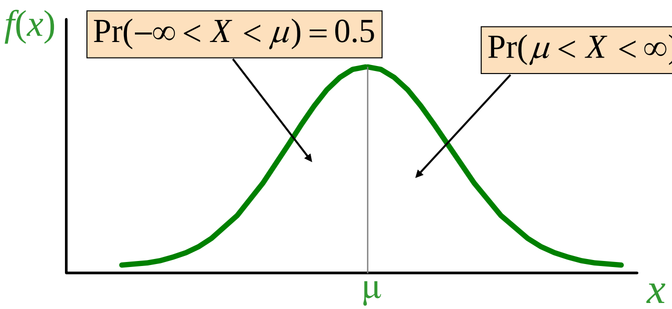


Finding Normal Probabilities – II

- The *total area under the curve is 1.0*, and the curve is symmetric, so half is above the mean, half is below

$\Pr(-\infty < X < \mu) = 0.5$

$\Pr(\mu < X < \infty) = 0.5$

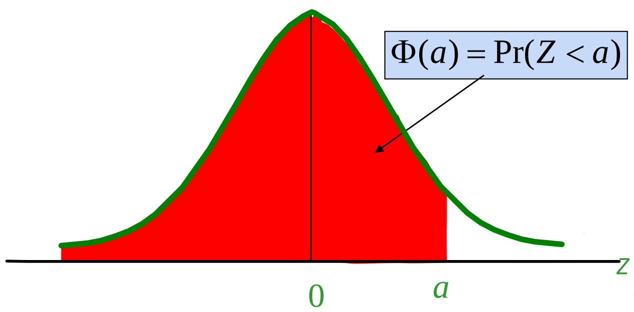


$\Pr(-\infty < X < \infty) = 1.0$

Finding Normal Probabilities – III

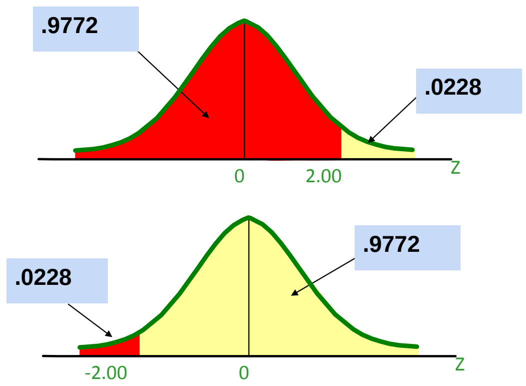
- Table with cumulative *standard normal distribution*: For a given Z-value a, the table shows $\Phi(a)$ (the area under the curve from negative infinity to a)

$\Phi(a) = \Pr(Z < a)$



Finding Normal Probabilities – IV

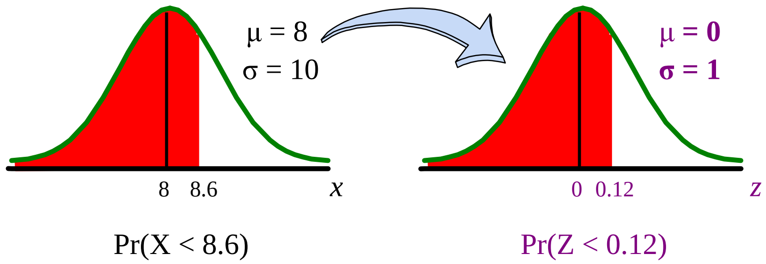
- **Example**: Suppose we are interested in $\Pr(Z < 2)$ – from the previous example. For negative $Z-$values, we use the fact that the distribution is symmetric to find the needed probability (e.g. $\Pr(Z < -2)$).



Finding Normal Probabilities – V

- **Example**: Suppose X is normal with mean 8.0 and standard deviation 5.0. Find $\Pr(X < 8.6)$.

$$Z = \frac{X-\mu}{\sigma} = \frac{8.6-8.0}{5.0} = 0.12;$$
$$\Phi(0.12) = 0.5478$$



Finding Normal Probabilities – VI

- **Example (Upper Tail Probabilities)**: Suppose X is normal with mean 8.0 and standard deviation 5.0. Find $\Pr(X > 8.6)$.

$$\begin{aligned}\Pr(X &> 8.6) = \Pr(Z > 0.12) = 1 - \Pr(Z \leq 0.12) \\ &= 1 - 0.5478 = 0.4522\end{aligned}$$

- **Example (Finding X for a Known Probability)** Suppose $X \sim N(8, 5^2)$. Find a X value so that only 20% of all values are below this X.
  1. Find the Z-value for the known probability
     $\Phi(.84) = .7995$, so a 20% area in the lower tail is consistent with a Z-value of -0.84.

Finding Normal Probabilities – VII

2. Convert to X-units using the formula

$$\begin{aligned}X &= \mu + Z\sigma \\ &= 8 + (-.84) \cdot 5 = 3.8.\end{aligned}$$

So 20% of the values from a distribution with mean 8 and standard deviation 5 are less than 3.80.

Joint and Marginal Probability Distributions – I

Joint Probability Functions

- Suppose that X and Y are discrete random variables. The ***joint probability function*** is

$$P(x, y) = \Pr(X = x \cap Y = y),$$

which is simply used to express the probability that X takes the specific value x and simultaneously Y takes the value y, as a function of x and y. This should satisfy:

1. $0 \leq P(x, y) \leq 1$ for all x, y.
2. $\sum_x \sum_y P(x, y) = 1$, where the sum is over all values (x, y) that are assigned nonzero probabilities.

Joint and Marginal Probability Distributions – II

Joint Probability Functions

- For any random variables X and Y (discrete or continuous), the ***joint (bivariate) distribution function*** $F(x, y)$ is

$$F(x, y) = \Pr(X \leq x \cap Y \leq y).$$

This defines the probability that simultaneously X is less than x and Y is less than y.

Joint and Marginal Probability Distributions

Marginal Probability Functions

- Let X and Y be jointly discrete random variables with probability function $P(x, y)$. Then the ***marginal probability functions*** of X and Y, respectively, are given by

$$P_x(x) = \sum_{\text{all } y} P(x, y) \qquad P_y(y) = \sum_{\text{all } x} P(x, y)$$

- Let X and Y be jointly discrete random variables with probability function $P(x, y)$. The ***cumulative marginal probability functions***, denoted $F_x(x_0)$ and $G_y(y_0)$, show the probability that X is less than or equal to x_0 and that Y is less than or equal to y_0 respectively

$$\begin{aligned} F_x(x_0) &= \Pr(X \leq x_0) = \sum_{x \leq x_0} P_x(x), \\ G_y(y_0) &= \Pr(Y \leq y_0) = \sum_{y \leq y_0} P_y(y). \end{aligned}$$

Conditional Probability Distributions

- If X and Y are jointly discrete random variables with joint probability function $P(x, y)$ and marginal probability functions $P_x(x)$ and $P_y(y)$, respectively, then the conditional discrete probability function of Y given X is

$$P(y|x) = \Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{P(x, y)}{P_x(x)},$$

provided that $P_x(x) > 0$. Similarly,

$$P(x|y) = \frac{P(x, y)}{P_y(y)}, \text{ provided that } P_y(x) > 0$$

Statistical Independence

- Let X have distribution function $F_x(x)$, Y have distribution function $F_y(y)$, and X and Y have a joint distribution function $F(x, y)$. Then X and Y are said to be ***independent*** if and only if

$$F(x, y) = F_x(x) \cdot F_y(y),$$

for every pair of real numbers (x, y).
- Alternatively, the two random variables X and Y are independent if the conditional distribution of Y given X does not depend on X:

$$\Pr(Y = y|X = x) = \Pr(Y = y).$$

- We also define Y to be **mean independent** of X when the conditional mean of Y given X equals the unconditional mean of Y:

$$E(Y = y|X = x) = E(Y = y).$$

Conditional Moments

- If X and Y are any two discrete random variables, the ***conditional expectation*** of Y given that $X = x$, is defined to be

$$\mu_{Y|X} = E(Y|X = x) = \sum_{\text{all } y} y \cdot P(y|x)$$

- If X and Y are any two discrete random variables, the ***conditional variance*** of Y given that $X = x$, is defined to be

$$\sigma^2_{Y|X} = E[(Y - \mu_{Y|X})^2|X = x] = \sum_{\text{all } y}(y - \mu_{Y|X})^2 \cdot P(y|x)$$

Joint and Marginal Distributions – I

Examples

- We are given the following data on the number of people attending AUEB this year.

	Subject of Study (Y)		
Sex (X)	***Economics*** (0)	***Finance*** (1)	***Systems*** (2)
Male (0)	40	10	30
Female (1)	30	20	70

1. What is the probability of selecting an individual that studies Finance?
2. What is the expected value of *Sex*?
3. What is the probability of choosing an individual that studies economics, given that it is a female?
4. Are *Sex* and *Subject* statistically independent?

Joint and Marginal Distributions – II

Examples

- **First step: Totals**

	Subject of Study (Y)			
Sex (X)	***Economics*** (0)	***Finance*** (1)	***Systems*** (2)	Total
Male (0)	40	10	30	**80**
Female (1)	30	20	70	**120**
Total	**70**	**30**	**100**	**200**

- **Second step: Probabilities**

	Subject of Study (Y)			
Sex (X)	***Economics*** (0)	***Finance*** (1)	***Systems*** (2)	Total
Male (0)	$40/200 = 0.20$	0.05	0.15	**0.40**
Female (1)	$30/200 = 0.15$	0.10	0.35	**0.60**
Total	$\mathbf{70/200 = 0.35}$	**0.15**	**0.50**	**1**

Joint and Marginal Distributions – III

Examples

- Answers:
  1. $\Pr(Y = 1) = 0.15$.
  2. $E(X) = 0 \cdot 0.4 + 1 \cdot 0.6 = 0.6$
  3. $\Pr(Y = 0|X = 1) = 0.15/0.6 = 0.25$
  4. $\Pr(X = 0 \cap Y = 0) = 0.20 \neq \Pr(X = 0) \cdot \Pr(Y = 0) = 0.4 \cdot 0.35 = 0.14$. So *Sex* and *Subject* are not statistically independent.
  - The conditional mean of Y given $X = 0$ is

$$\begin{aligned} &E(Y|X = 0) \\ &= \Pr(Y = 0|X = 0) \cdot 0 + \Pr(Y = 1|X = 0) \cdot 1 + \Pr(Y = 2|X = 0) \cdot 2 \\ &= \frac{0.20}{0.4} \cdot 0 + \frac{0.05}{0.4} \cdot 1 + \frac{0.15}{0.4} \cdot 2 = 0.875 \end{aligned}$$

Joint and Marginal Distributions – IV

Examples

- The conditional mean of Y given $X = 1$ is

$$\begin{aligned} &E(Y|X = 1) \\ &= \Pr(Y = 0|X = 1) \cdot 0 + \Pr(Y = 1|X = 1) \cdot 1 + \Pr(Y = 2|X = 1) \cdot 2 \\ &= \frac{0.15}{0.6} \cdot 0 + \frac{0.10}{0.6} \cdot 1 + \frac{0.35}{0.6} \cdot 2 = 0.80 \end{aligned}$$

Covariance, Correlation and Independence – I

Definition (Covariance)

If X and Y are random variables with means μ_x and μ_y, respectively, the ***covariance*** of X and Y is

$$\sigma_{XY} \equiv \mathrm{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)].$$

- This can be found as

$$\mathrm{Cov}(X, Y) = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_x)(y - \mu_y) \cdot P(x, y),$$

and an equivalent expression is

$$\mathrm{Cov}(X, Y) = E[XY] - \mu_x\mu_y = \sum_{\text{all } x} \sum_{\text{all } y} xy \cdot P(x, y) - \mu_x\mu_y.$$

Covariance, Correlation and Independence – II

- The *covariance* measures the strength of the linear relationship between two variables.
- If two random variables are statistically independent, the covariance between them is 0. The converse is **not** necessarily true.

Covariance, Correlation and Independence – III

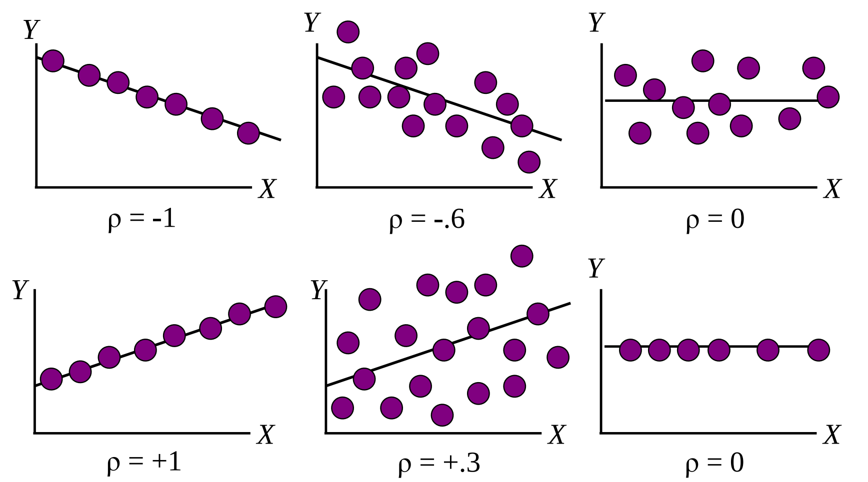
Definition (Correlation)

The correlation between X and Y is

$$\rho \equiv \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

- $\rho = 0 \Rightarrow$ no linear relationship between X and Y.
- $\rho > 0 \Rightarrow$ positive linear relationship between X and Y.
  - when X is high (low) then Y is likely to be high (low)
  - $\rho = +1 \Rightarrow$ perfect positive linear dependency
- $\rho < 0 \Rightarrow$ negative linear relationship between X and Y.
  - when X is high (low) then Y is likely to be low (high)
  - $\rho = -1 \Rightarrow$ perfect negative linear dependency

Covariance, Correlation and Independence – IV



Moments of Linear Combinations – I

- Let X and Y be two random variables with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2 and covariance $\mathrm{Cov}(X,Y)$. Take a linear combination of X and Y :

$$W = aX + bY.$$

Then,

$$\mathrm{E}(W) = \mathrm{E}(aX + bY) = a\mu_X + b\mu_Y, \text{ and}$$

$$\mathrm{Var}(W) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\mathrm{Cov}(X,Y),$$

or using the correlation

$$\mathrm{Var}(W) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\mathrm{Corr}(X,Y)\sigma_X\sigma_Y$$

Moments of Linear Combinations – II

Example

If $a = 1$ and $b = -1$, $W = X - Y$ and

$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(X - Y) = \mu_X - \mu_Y \\ \mathrm{Var}(W) &= \sigma_X^2 + \sigma_Y^2 - 2\mathrm{Cov}(X,Y) \\ &= \sigma_X^2 + \sigma_Y^2 - 2\mathrm{Corr}(X,Y)\sigma_X\sigma_Y \end{aligned}$$

Moments of Linear Combinations

Example 1: Normally Distributed Random Variables

- Two tasks must be performed by the same worker.
  - X = minutes to complete task 1; $\mu_X = 20, \sigma_X = 5$;
  - Y = minutes to complete task 2; $\mu_Y = 30, \sigma_Y = 8$;
  - X and Y are normally distributed and independent...
- ★ What is the mean and standard deviation of the time to complete both tasks?
- $W = X + Y$ (total time to complete both tasks). So

$$
\begin{aligned}
\mathrm{E}(W) &= \mu_X + \mu_Y = 20 + 30 = 50 \\
\mathrm{Var}(W) &= \sigma_X^2 + \sigma_Y^2 + \underbrace{2\mathrm{Cov}(X, Y)}_{=0,\ \text{independence}} = 5^2 + 8^2 = 89 \\
\Rightarrow \sigma_W &= \sqrt{89} \simeq 9.43
\end{aligned}
$$

Linear Combinations Random Variables – I

Example 2: Portfolio Value

- The return per $\$1,000$ for two types of investments is given below

State of Economy		Investment Funds	
Prob	**Economic condition**	***Passive*** X	***Aggressive*** Y
0.2	Recession	$-\$25$	$-\$200$
0.5	Stable Economy	$+\$50$	$+\$60$
0.3	Growing Economy	$+\$100$	$+\$350$

- Suppose 40% of the portfolio (P) is in Investment X and 60% is in Investment Y. Calculate the portfolio return and risk.
  - Mean return for each fund investment

$$
\begin{aligned}
\mathrm{E}(X) &= \mu_X = (-25)(.2) + (50)(.5) + (100)(.3) = 50 \\
\mathrm{E}(Y) &= \mu_Y = (-200)(.2) + (60)(.5) + (350)(.3) = 95
\end{aligned}
$$

Linear Combinations Random Variables – II

Example 2: Portfolio Value

- Standard deviations for each fund investment

$$
\begin{aligned}
\sigma_X &= \sqrt{(-25-50)^2(.2) + (50-50)^2(.5) + (100-50)^2(.3)} \\
&= 43.30 \\
\sigma_Y &= \sqrt{(-200-95)^2(.2) + (60-95)^2(.5) + (350-95)^2(.3)} \\
&= 193.71
\end{aligned}
$$

- The covariance between the two fund investments is

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= (-25-50)(-200-95)(.2) \\
&\quad + (50-50)(60-95)(.5) \\
&\quad + (100-50)(350-95)(.3) \\
&= 8250
\end{aligned}
$$

Linear Combinations Random Variables – III

Example 2: Portfolio Value

- So

$$
\begin{aligned}
\mathrm{E}(P) &= 0.4(50) + 0.6(95) = 77 \\
\sigma_P &= \sqrt{(.4)^2(43.30)^2 + (.6)^2(193.71)^2 + 2(.4)(.6)8250} \\
&= 133.04
\end{aligned}
$$

The t-Distribution – I

- Let two independent random variables $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$.[1] If Z and Y are independent, then

$$W = \frac{Z}{\sqrt{Y/n}} \sim t(n)$$

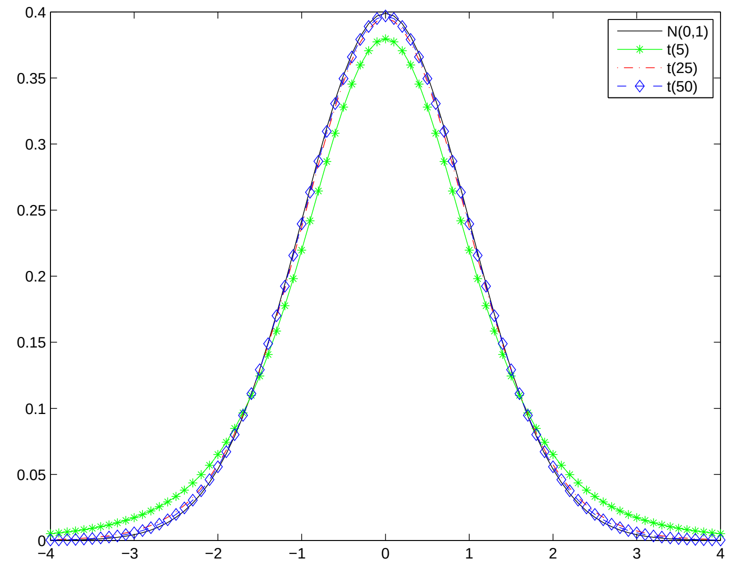
  - The PDF of t has only one parameter, n, is always positive and symmetric around zero.
  - Moreover it holds that

$$\mathrm{E}(W) = 0 \text{ for } n > 1; \quad \mathrm{Var}(W) = \frac{n}{n-2} \text{ for } n > 2$$

and for n large enough: $W \underset{n\to\infty}{\sim} N(0,1)$

[1] Let $Z_1, Z_2, \ldots, Z_n$ be independent r.v.s and $Z_i \sim N(0,1)$. Then $\Upsilon = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$.

The t-Distribution – II



Annex: Normal Approximation of Binomial – I

- Recall the binomial distribution, where we have n *independent trials* and the probability of success on any given trial = p.
- Let X be a binomial random variable ($X_i = 1$ if the ith trial is "success"):

$$\begin{aligned} \mathrm{E}(X) &= \mu = np \\ \mathrm{Var}(X) &= \sigma^2 = np(1-p) \end{aligned}$$

  - The shape of the binomial distribution is approximately normal if n is large

Annex: Normal Approximation of Binomial – II

- The normal is a good approximation to the binomial when $np(1-p) > 5$ (check that $np > 5$ and $n(1-p) > 5$ to be on the safe side). That is

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}}.$$

- For instance, let X be the number of successes from n independent trials, each with probability of success p. Then

$$\Pr(a < X < b) = \Pr\left(\frac{a - np}{\sqrt{np(1-p)}} < Z < \frac{b - np}{\sqrt{np(1-p)}}\right)$$

Annex: Normal Approximation of Binomial – III

- **Example**: 40% of all voters support ballot proposition A. What is the probability that between 76 and 80 voters indicate support in a sample of n = 200?

$$\begin{aligned} \mathrm{E}(X) &= \mu = np = 200(0.40) = 80 \\ \mathrm{Var}(X) &= np(1-p) = 200(0.40)(1-0.40) = 48 \end{aligned}$$

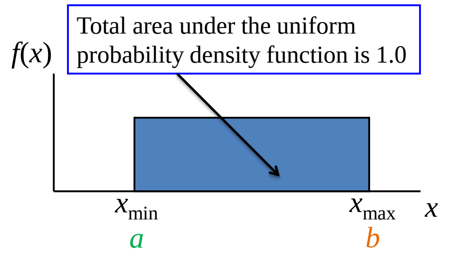
So

$$\Pr(76 < X < 80) = \Pr\left(\frac{76-80}{\sqrt{48}} < Z < \frac{80-80}{\sqrt{48}}\right)$$
$$= \Pr(-0.58 < Z < 0)$$
$$= \Phi(0) - \Phi(-0.58)$$
$$= 0.500 - 0.2810 = 0.219$$

Annex: Uniform Distribution – I

- The *uniform distribution* is a probability distribution that has *equal probabilities* for all possible outcomes of the random variable (where $x_{\min} = a$ and $x_{\max} = b$)

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \; ; \; F(x) \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & x \ge b \end{cases}$$



Annex: Uniform Distribution – II

- Moments uniform distribution

$$\mu = \frac{a+b}{2}; \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

- **Example**: Uniform probability distribution over the range $2 \le x \le 6$. Then

$$f(x) = \frac{1}{6-2} = 0.25 \text{ for } 2 \le x \le 6$$

and

$$\begin{aligned} \mathrm{E}(X) &= \mu = \frac{a+b}{2} = \frac{2+6}{2} = 4 \\ \mathrm{Var}(X) &= \sigma^2 = \frac{(b-a)^2}{12} = \frac{(6-2)^2}{12} = 1.333 \end{aligned}$$

Annex: The χ^2 Distribution – I

- Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables and $Z_i \sim N(0,1)$. Then

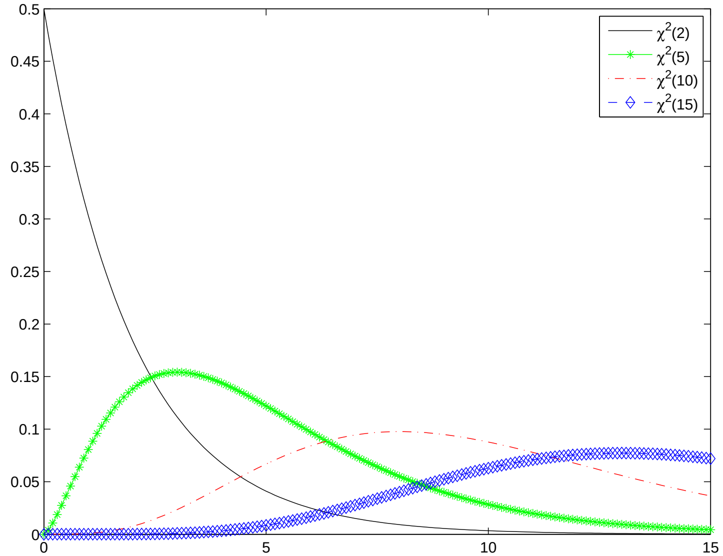
$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

  - The PDF of χ^2 has only one parameter, n, is always positive and right asymmetric.
  - Moreover it holds that

$$\begin{aligned} \mathrm{E}(X) &= n; \text{ and} \\ \mathrm{Var}(X) &= 2n \end{aligned}$$

for $n \ge 2$.

Annex: The χ^2 Distribution – II



Annex: The F Distribution – I

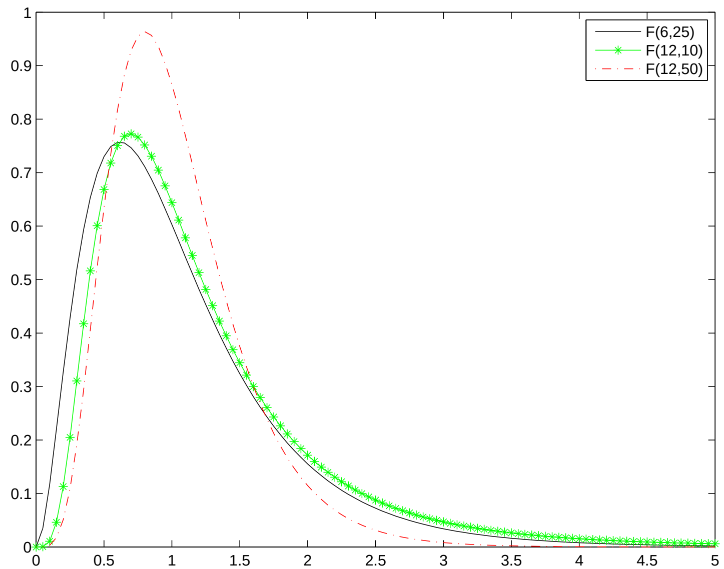
- Let X and Y be two independent random variables, that are distributed as $\chi^2 : X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$. Then

$$W = \frac{X/n}{Y/m} \sim F(n, m)$$

  - The PDF of F has two parameters, n and m (the degrees of freedom of the numerator and the denominator); it is positive and right asymmetric.
  - Moreover it holds that if $W \sim F(n, m)$

$$\mathrm{E}(W) = \frac{m}{1-m}; \text{ for } m > 2.$$

Annex: The F Distribution – II



Statistics for Business

Sampling Distributions, Interval Estimation and Hypothesis Tests.

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: July 15, 2015. **This Draft**: August 28, 2023.

Lecture Outline

- Simple random sampling
- Distribution of the sample average
- Large sample approximation to the distribution of the sample mean
  - Law of Large Numbers
  - Central Limit Theorem
- Estimation of the population mean
  - Unbiasedness
  - Consistency
  - Efficiency
- Hypothesis test concerning the population mean
- Confidence intervals for the population mean
  - Using the t-statistic when n is small
- Comparing means from different populations

Sampling

- A ***population*** is a collection of all the elements of interest, while a ***sample*** is a subset of the population.
- The reason we select a sample is to collect data to answer a research question about a population.
- The sample results provide only **estimates** of the values of the population characteristics. With *proper sampling methods*, the sample results can provide "good" estimates of the population characteristics.
- A ***random sample*** from an infinite population is a sample selected such that the following conditions are satisfied:
  - Each element selected comes from the population of interest.
  - Each element is selected *independently*.
  - ★ If the population is finite, then we sample with replacement...

Simple Random Sampling – I

- ***Simple random sampling*** means that n objects are drawn randomly from a population and each object is equally likely to be drawn
- Let $Y_1, Y_2, \ldots, Y_n$ denote the 1st to the n th randomly drawn object. Under simple random sampling
  - The marginal probability distribution of Y_i is the same for all $i = 1, 2, \ldots, n$ and equals the population distribution of Y.
  - ⋆ because $Y_1, Y_2, \ldots, Y_n$ are drawn randomly from the **same** population.
  - Y_1 is distributed independently from $Y_2, \ldots, Y_n$. knowing the value of Y_i does not provide information on Y_j for $i \neq j$
- When $Y_1, Y_2, \ldots, Y_n$ are drawn from the same population and are independently distributed, they are said to be ***I.I.D. random variables***

Simple Random Sampling – II

Example

- Let G be the gender of an individual ($G = 1$ if female, $G = 0$ if male)
- G is a Bernoulli r.v. with $\mathrm{E}(G) = \mu_G = \Pr(G = 1) = 0.5$
- Suppose we take the population register and randomly draw a sample of size n
  - The probability distribution of G_i is a Bernoulli with mean 0.5
  - G_1 is distributed independently from $G_2, \ldots, G_n$
- Suppose we draw a random sample of individuals entering the building of the accounting department
  - This is not a sample obtained by simple random sampling and $G_1, G_2, \ldots, G_n$ are not i.i.d
  - Men are more likely to enter the building of the accounting department!

The Sampling Distribution of the Sample Average – I

- The ***sample average*** $\bar{Y}$ of a randomly drawn sample is a random variable with a probability distribution called the ***sampling distribution***

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  - The individuals in the sample are drawn at random.
  - Thus the values of $(Y_1, Y_2, \cdots, Y_n)$ are random
  - Thus functions of $(Y_1, Y_2, \cdots, Y_n)$, such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value
  - The distribution of over different possible samples of size n is called the ***sampling distribution*** of $\bar{Y}$.
  - The mean and variance of are the mean and variance of its sampling distribution, $\mathrm{E}(\bar{Y})$ and $\mathrm{Var}(\bar{Y})$.
  - The concept of the sampling distribution underpins all of statistics/econometrics.

The Sampling Distribution of the Sample Average – II

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Suppose that $Y_1, Y_2, \ldots, Y_n$ are *I.I.D.* and the mean & variance of the population distribution of Y are respectively μ_Y and σ_Y^2
  - The mean of (the sampling distribution of) $\bar{Y}$ is

$$\mathrm{E}(\bar{Y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i) = \frac{1}{n}n\mathrm{E}(Y) = \mu_Y$$

  - The variance of (the sampling distribution of) $\bar{Y}$ is

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) + 2\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathrm{Cov}(Y_i, Y_j) \\ &= \frac{1}{n^2}n\mathrm{Var}(Y) + 0 = \frac{1}{n}\mathrm{Var}(Y) = \frac{\sigma_Y^2}{n}\end{aligned}$$

The Sampling Distribution of the Sample Average – III

Example

- Let G be the gender of an individual ($G = 1$ if female, $G = 0$ if male)
- The mean of the population distribution of G is

$$\mathrm{E}(G) = \mu_G = \Pr(G = 1) = p = 0.5$$

- The variance of the population distribution of G is

$$\mathrm{Var}(G) = \sigma_G^2 = p(1 - p) = 0.5(1 - 0.5) = 0.25$$

- The mean and variance of the average gender (proportion of women) $\bar{G}$ in a random sample with $n = 10$ are

$$\begin{aligned}\mathrm{E}(\bar{G}) &= \mu_G = 0.5 \\ \mathrm{Var}(\bar{G}) &= \frac{1}{n}\sigma_G^2 = \frac{1}{10}0.25 = 0.025\end{aligned}$$

The Finite-Sample Distribution of the Sample Average

- The ***finite sample distribution*** is the sampling distribution that exactly describes the distribution of $\bar{Y}$ for any sample size n.
- In general the exact sampling distribution of $\bar{Y}$ is complicated and depends on the population distribution of Y.
- A special case is when $Y_1, Y_2, \ldots, Y_n$ are *IID* draws from the $N(\mu_Y, \sigma_Y^2)$, because in this case

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

The Sampling Distribution of the Average Gender $\bar{G}$

- Suppose G takes on 0 or 1 (a Bernoulli random variable) with the probability distribution

$$\Pr(G=0)=p=0.5, \quad \Pr(G=1)=1-p=0.5$$

- As we discussed above:

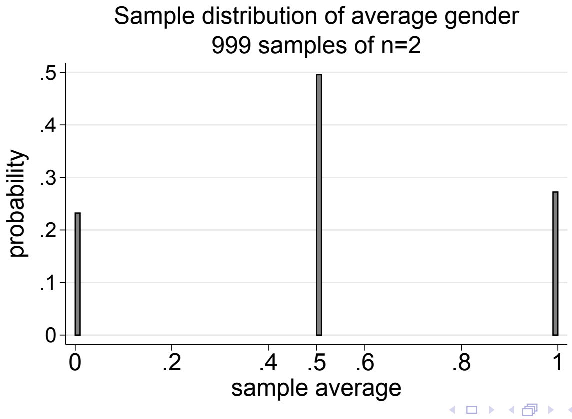
$$\begin{aligned} \mathrm{E}(G) &= \mu_G = \Pr(G=1) = p = 0.5 \\ \mathrm{Var}(G) &= \sigma_G^2 = p(1-p) = 0.5(1-0.5) = 0.25 \end{aligned}$$

- The sampling distribution of $\bar{G}$ depends on n.
- Consider $n=2$. The sampling distribution of $\bar{G}$ is
  - $\Pr(\bar{G}=0)=0.5^2=0.25$
  - $\Pr(\bar{G}=1/2)=2\times 0.5\times(1-0.5)=0.5$
  - $\Pr(\bar{G}=1)=(1-0.5)^2=0.25$

The Finite-Sample Distribution of the Average Gender $\bar{G}$

- Suppose we draw 999 samples of $n=2$:

Sample 1			Sample 1			Sample 3			$\cdots$	Sample 999		
G_1	G_2	$\bar{G}$	G_1	G_2	$\bar{G}$	G_1	G_2	$\bar{G}$		G_1	G_2	$\bar{G}$
1	0	0.5	1	1	1	0	1	0.5		0	0	0



The Asymptotic Distribution of the Sample Average $\bar{Y}$

- Given that the exact sampling distribution of $\bar{Y}$ is complicated and given that we generally use large samples in statistics/econometrics we will often use an approximation of the sample distribution that relies on the sample being large
- The ***asymptotic distribution*** or ***large-sample distribution*** is the approximate sampling distribution of $\bar{Y}$ if the sample size becomes very large: $n\to\infty$.
- We will use two concepts to approximate the large-sample distribution of the sample average
  - The law of large numbers.
  - The central limit theorem.

The Law of Large Numbers (LLN)

Definition (Law of Large Numbers)

Suppose that

1. $Y_i, i=1,\ldots,n$ are independently and identically distributed with $\mathrm{E}(Y_i)=\mu_Y$; and
2. large outliers are unlikely i.e. $\mathrm{Var}(Y_i)=\sigma_Y^2<+\infty$.

Then $\bar{Y}$ will be near μ_Y with very high probability when n is very large $(n\to\infty)$

$$\bar{Y}\xrightarrow{p}\mu_Y.$$

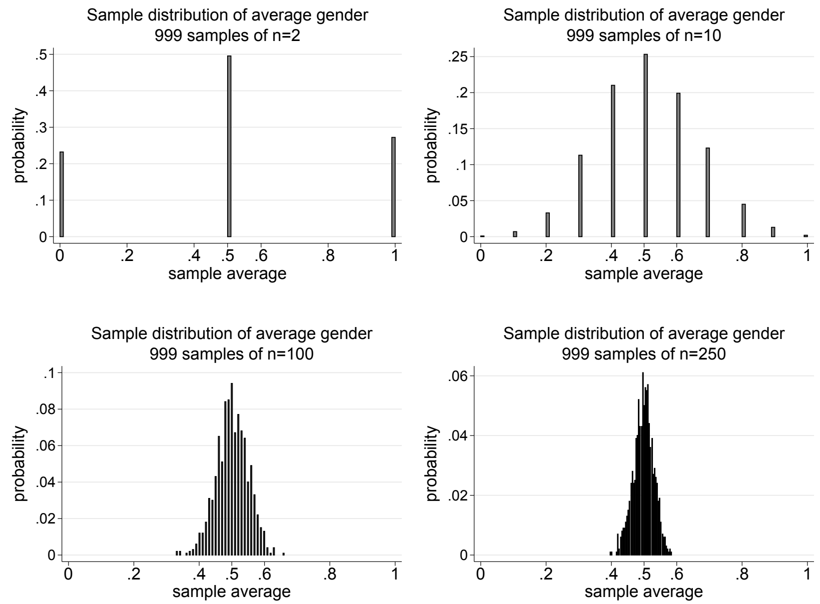
We also say that the sequence of random variables $\{Y_n\}$ converges in probability to the μ_Y, if for every $\varepsilon>0$

$$\lim_{n\to\infty}\Pr(|\bar{Y}_n-\mu_Y|>\varepsilon)=0.$$

We also denote this by $\mathrm{plim}(Y_n)=\mu_Y$

The Law of Large Numbers (LLN)

Example: Gender $G \sim Bernoulli(0.5, 0.25)$



The Central Limit Theorem (CLT)

Definition (Central Limit Theorem)

Suppose that

1. $Y_i, i = 1, \ldots, n$ are independently and identically distributed with $\mathrm{E}(Y_i) = \mu_Y$; and
2. large outliers are unlikely i.e. $\mathrm{Var}(Y_i) = \sigma_Y^2$ with $0 < \sigma_Y^2 < +\infty$.

Then the distribution of the sample average $\bar{Y}$ will be approximately normal as n becomes very large ($n \to \infty$)

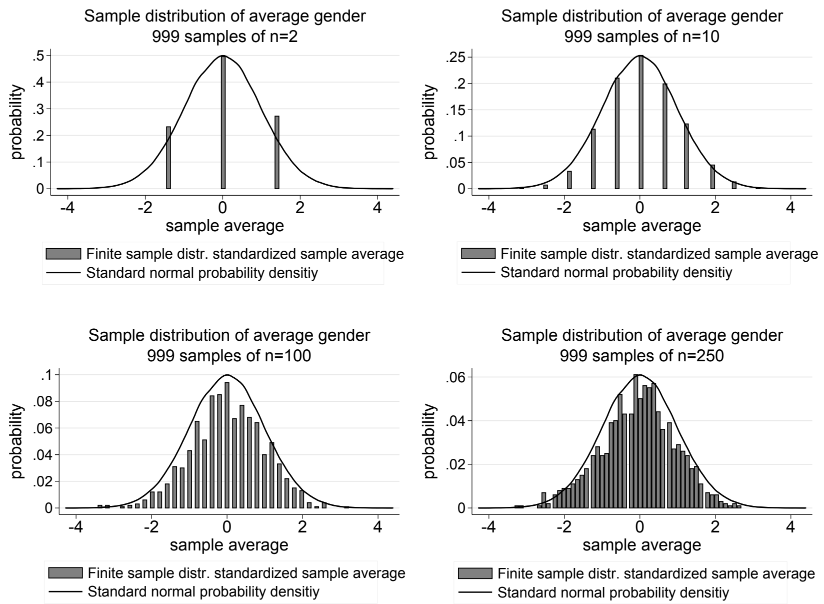
$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right).$$

The distribution of the the standardized sample average is approximately standard normal for $n \to \infty$

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$$

The Central Limit Theorem (CLT)

Example: Gender $G \sim Bernoulli(0.5, 0.25)$



The Central Limit Theorem (CLT)

- How good is the large-sample approximation?
- ⋆ If $Y_i \sim N(\mu_Y, \sigma_Y^2)$ the approximation is perfect.
- ⋆ If Y_i is not normally distributed the quality of the approximation depends on how close n is to infinity (how large n is)
- ⋆ For $n \geq 100$ the normal approximation to the distribution of $\bar{Y}$ is typically very good for a wide variety of population distributions.

Estimators and Estimates

Definition

An ***estimator*** is a function of a sample of data to be drawn randomly from a population.

- An estimator is a random variable because of randomness in drawing the sample. Typically used estimators

  Sample Average: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, Sample variance: $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

  Using a particular sample $y_1, y_2, ..., y_n$ we obtain

  $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

  which are ***point estimates***. These are the numerical value of an estimator when it is actually computed using a specific sample.

Estimation of the Population Mean – I

- Suppose we want to know the mean value of Y (μ_Y) in a population, for example
  - The mean wage of college graduates.
  - The mean level of education in Greece.
  - The mean probability of passing the statistics exam.
- Suppose we draw a random sample of size n with $Y_1, Y_2, ..., Y_n$ being *IID*
- Possible estimators of μ_Y are:
  - The sample average: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
  - The first observation: Y_1
  - The weighted average: $\tilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + ... + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right)$.
- To determine which of the estimators, $\bar{Y}$, Y_1 or $\tilde{Y}$ is the best estimator of μ_Y we consider 3 properties.
- Let $\hat{\mu}_Y$ be an estimator of the population mean μ_Y

Estimation of the Population Mean – II

1. **Unbiasedness**: The mean of the sampling distribution of $\hat{\mu}_Y$ equals μ_Y

   $$\mathrm{E}(\hat{\mu}_Y) = \mu_Y.$$

2. **Consistency**: The probability that $\hat{\mu}_Y$ is within a very small interval of μ_Y approaches 1 if $n \to \infty$

   $$\hat{\mu}_Y \xrightarrow{p} \mu_Y \text{ or } \Pr(|\hat{\mu}_Y - \mu_Y| < \varepsilon) = 1$$

3. **Efficiency**: If the variance of the sampling distribution of $\hat{\mu}_Y$ is smaller than that of some other estimator $\tilde{\mu}_Y$, $\hat{\mu}_Y$ is more efficient

   $$\mathrm{Var}(\hat{\mu}_Y) \leq \mathrm{Var}(\tilde{\mu}_Y)$$

Estimating Mean Wages – I

- Suppose we are interested in the mean wages (pre tax) μ_W of individuals with a Ph.D. in economics/finance in Europe (true mean $\mu_w = 60K$). We draw the following sample ($n = 10$) by simple random sampling

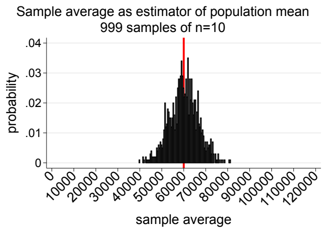
i	1	2	3	4	5
W_i	47281.92	70781.94	55174.46	49096.05	67424.82

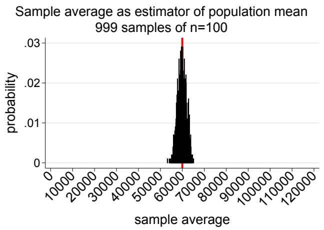
i	6	7	8	9	10
W_i	39252.85	78815.33	46750.78	46587.89	25015.71

- The 3 estimators give the following estimates:
  - $\bar{W} = \frac{1}{10}\sum_{i=1}^{10} W_i = 52618.18$
  - $W_1 = 47281.92$
  - $\tilde{W} = \frac{1}{10}\left(\frac{1}{2}W_1 + \frac{3}{2}W_2 + ... + \frac{1}{2}W_9 + \frac{3}{2}W_{10}\right) = 49398.82$
- **Unbiasedness**: All 3 proposed estimators are unbiased

Estimating Mean Wages – II

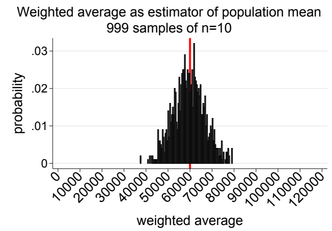
- **Consistency**:
  - By the law of large numbers $\bar{W} \xrightarrow{p} \mu_W$ which implies that the probability that $\bar{W}$ is within a very small interval of μ_W approaches 1 if $n \to \infty$

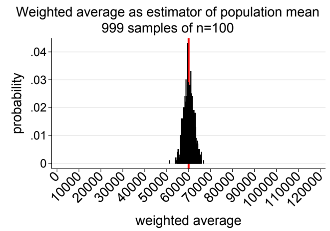




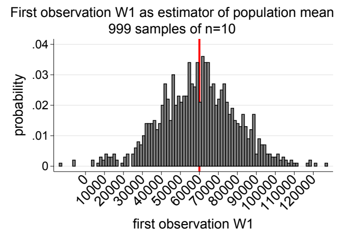
Estimating Mean Wages – III

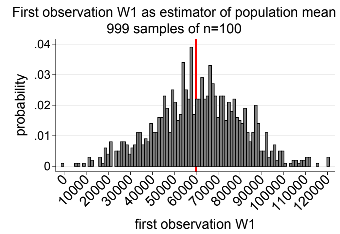
- $\tilde{W} = \frac{1}{n}\left(\frac{1}{2}W_1 + \frac{3}{2}W_2 + \ldots + \frac{1}{2}W_{n-1} + \frac{3}{2}W_n\right)$ can also be shown to be consistent





- However W_1 is not a consistent estimator of μ_W.





Estimating Mean Wages – IV

- **Efficiency**: We have that
  - $\text{Var}(\bar{W}) = \frac{1}{n}\sigma_W^2$
  - $\text{Var}(W_1) = \sigma_W^2$
  - $\text{Var}(\tilde{W}) = 1.25\frac{1}{n}\sigma_W^2$
  - So for any $n \geq 2$, $\bar{W}$ is more efficient than W_1 and $\tilde{W}$.
- In fact **$\bar{Y}$ is the Best Linear Unbiased Estimator (BLUE)**: it is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, Y_2, \ldots, Y_n$

⋆ Let $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} \alpha_i Y_i$ be an unbiased estimator of μ_Y with α_i nonrandom constants. Then $\bar{Y}$ is more efficient than $\hat{\mu}_Y$

$$\text{Var}(\bar{Y}) \leq \text{Var}(\hat{\mu}_Y)$$

Hypothesis Tests

Consider the following questions:

- Is the mean monthly wage of Ph.D. graduates equal to 60000 euros?
- Is the mean level of education in Greece equal to 12 years?
- Is the mean probability of passing the stats exam equal to 1?

These questions involve the population mean taking on a specific value $\mu_{Y,0}$. Answering these questions implies using data to compare a ***null hypothesis*** (a tentative assumption about the population mean parameter)

$$H_0 : \text{E}(Y) = \mu_{Y,0}$$

to an ***alternative hypothesis*** (the opposite of what is stated in the H_0)

$$H_1 : \text{E}(Y) \neq \mu_{Y,0}$$

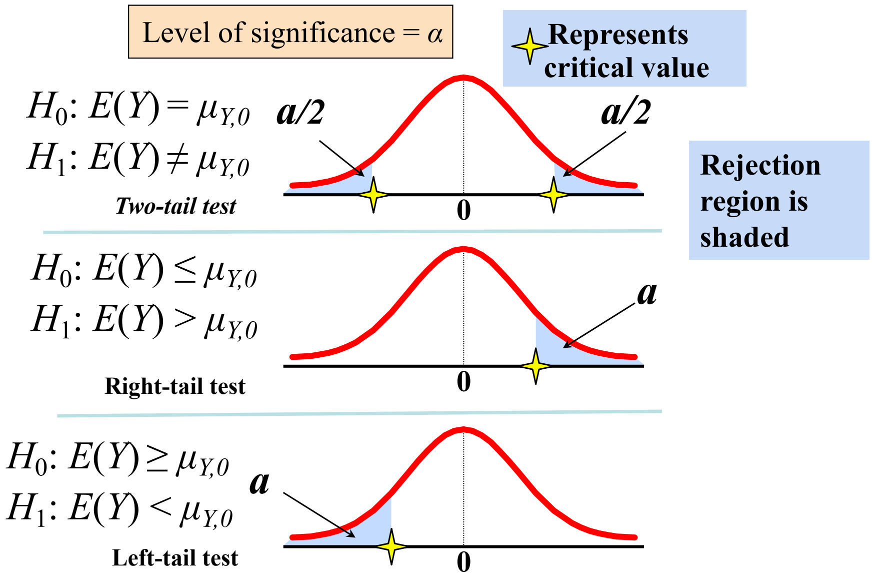
- Alternative Hypothesis as a Research Hypothesis
  - *Example*: A new sales force bonus plan is developed in an attempt to increase sales.
  - **Alternative Hypothesis**: The new bonus plan increase sales.
  - **Null Hypothesis**: The new bonus plan does not increase sales.

Hypothesis Tests: Terminology

- The **hypothesis testing problem** (for the mean): make a provisional decision, based on the evidence at hand, whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test
  - $H_0 : \mathrm{E}(Y) \leq \mu_{Y,0}$ vs. $H_1 : \mathrm{E}(Y) > \mu_{Y,0}$ (1-sided, $>$)
  - $H_0 : \mathrm{E}(Y) \geq \mu_{Y,0}$ vs. $H_1 : \mathrm{E}(Y) < \mu_{Y,0}$ (1-sided, $<$)
  - $H_0 : \mathrm{E}(Y) = \mu_{Y,0}$ vs. $H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$ (2-sided)
- p-value = probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.
- The **significance level** of a test (α) is a pre-specified probability of incorrectly rejecting the null, when the null is true. Typical values are 0.01 (1%), 0.05 (5%), or 0.10 (10%).
  - It is selected by the researcher at the beginning, and determines the ***critical value(s)*** of the test.
  - If the test-statistic falls outside the non-rejection region, we reject H_0.

Hypothesis Tests

The Testing Process and Rejections



Hypothesis Testing using p-values

- The p-value is the probability, computed using the test statistic, that measures the support (or lack of support) provided by the sample for the null hypothesis
  - If the p-value is less than or equal to the level of significance α, the value of the test statistic is in the rejection region.
  - Reject H_0 if the p-value $< \alpha$.
  - See also Annex
- **Rules of thumb**
  - If p-value is less than .01, there is overwhelming evidence to conclude H_0 is false.
  - If p-value is between .01 and .05, there is strong evidence to conclude H_0 is false.
  - If p-value is between .05 and .10, there is weak evidence to conclude H_0 is false.
  - If p-value is greater than .10, there is insufficient evidence to conclude H_0 is false.

Hypothesis Test for the Mean with σ_Y^2 known – I

Decision Rules

- The test statistic employed is obtained by converting the sample result ($\bar{y}$) to a z-value

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}$$

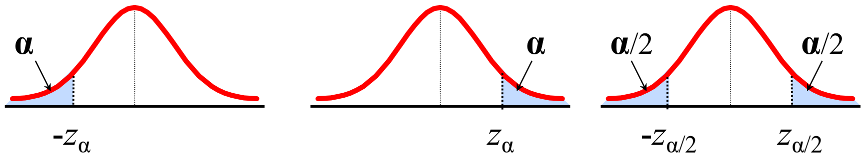
$H_0 : \mathrm{E}(Y) \geq \mu_{Y,0}$ $H_1 : \mathrm{E}(Y) < \mu_{Y,0}$	$H_0 : \mathrm{E}(Y) \leq \mu_{Y,0}$ $H_1 : \mathrm{E}(Y) > \mu_{Y,0}$	$H_0 : \mathrm{E}(Y) = \mu_{Y,0}$ $H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$
Lower-tail	Upper-tail	Two-tailed
Reject H_0 if $z < z_\alpha$	Reject H_0 if $z > z_\alpha$	Reject H_0 if $z < -z_{\alpha/2}$ or if $z > z_{\alpha/2}$

Hypothesis Test for the Mean with σ_Y^2 **known** – II

Decision Rules

Hypothesis Tests for $E(Y)$ $\quad z = \dfrac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} = \dfrac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(Y) \geq \mu_0$ H_1: $E(Y) < \mu_0$	H_0: $E(Y) \leq \mu_{Y,0}$ H_1: $E(Y) > \mu_{Y,0}$	H_0: $E(Y) = \mu_{Y,0}$ H_1: $E(Y) \neq \mu_{Y,0}$



Reject H_0 if $z < -z_\alpha$	Reject H_0 if $z > z_\alpha$	Reject H_0 if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Hypothesis Test for the Mean (σ^2 **known**) – I

Examples

- **Example 1**. A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over \$52 per month. The company wishes to test this claim. Assume $\sigma = 10\$$ is known and let $\alpha = 0.10$. Suppose a sample of 64 persons is taken, and it is found that the average bill \$53.1.
  - Form the hypothesis to be tested

$$
\begin{array}{ll}
H_0 : \mathrm{E}(Y) \leq 52 & \text{the } \textit{mean} \textbf{ is not} \text{ over \$52 per month} \\
H_1 : \mathrm{E}(Y) > 52 & \text{the } \textit{mean} \textbf{ is} \text{ over \$52 per month}
\end{array}
$$

  - For $\alpha = 0.10$, $z_{0.10} = 1.28$, so we would reject H_0 if $z > 1.28$.
  - We have $n = 64$ and $\bar{y} = 53.1$, so the test statistic is

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}} = \frac{53.1 - 52}{10/\sqrt{64}} = 0.88 < z_{0.10} = 1.28$$

Hence H_0 cannot be rejected.

Hypothesis Test for the Mean (σ^2 **known**) – II

Examples

- **Example 2**. We would like to test the claim that the true mean # of TV sets in EU homes is equal to 3 (assuming $\sigma_Y = 0.8$ known). For this purpose a sample of 100 homes is selected, and the average number of TV sets is 2.84. Test the above hypothesis using $\alpha = 0.05$.
  - Form the hypothesis to be tested

$$
\begin{array}{ll}
H_0 : \mathrm{E}(Y) = 3 & \text{the } \textit{mean} \text{ \# } \textbf{is} \text{ 3 TV sets per home} \\
H_1 : \mathrm{E}(Y) \neq 3 & \text{the } \textit{mean} \textbf{ is not} \text{ 3 TV sets per home}
\end{array}
$$

  - For $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$, so we would reject H_0 if $|z| > 1.96$.

Hypothesis Test for the Mean (σ^2 **known**) – III

Examples

- We have $n = 100$ and $\bar{y} = 2.84$, so the test statistic is

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}} = \frac{2.84 - 3}{0.8/\sqrt{100}} = \frac{-0.16}{0.08} = -2 < -z_{0.025} = -1.96$$

or $|z| = 2 > 1.96$, Hence H_0 is rejected. We **conclude** that there is sufficient evidence that the mean number of TVs in EU homes is not equal to 3.

 Hypothesis Tests for the Population Mean

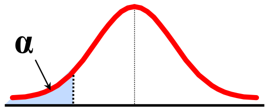
Test for the Mean with σ_Y^2 unknown but $n \to \infty$

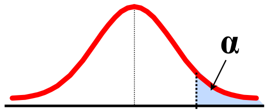
Decision Rules

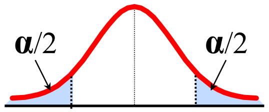
- Since $S_Y^2 \xrightarrow{p} \sigma_Y^2$, compute the standard error of $\bar{Y}$, $SE(\bar{Y}) = s_Y/\sqrt{n}$ and construct a t-ratio.

Hypothesis Tests for $E(Y)$ $\quad t = \dfrac{\bar{Y} - \mu_{Y,0}}{\mathrm{SE}(\bar{Y})} = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(Y) \geq \mu_0$ H_1: $E(Y) < \mu_0$	H_0: $E(Y) \leq \mu_{Y,0}$ H_1: $E(Y) > \mu_{Y,0}$	H_0: $E(Y) = \mu_{Y,0}$ H_1: $E(Y) \neq \mu_{Y,0}$







$-z_\alpha$	z_α	$-z_{\alpha/2}$ $\quad z_{\alpha/2}$
Reject H_0 if $t < -z_\alpha$	Reject H_0 if $t > z_\alpha$	Reject H_0 if $t < -z_{\alpha/2}$ or $t > z_{\alpha/2}$

 Hypothesis Tests for the Population Mean

Test for the Mean with σ_Y^2 unknown but $n \to \infty$

Example

- Suppose we would like to test

$$H_0 : \mathrm{E}(W) = 60000, \qquad H_1 : \mathrm{E}(W) \neq 60000,$$

  using a sample of 250 individuals with a Ph.D. degree at the 5% significance level.
- We perform the following steps:
  1. $\bar{W} = \frac{1}{n}\sum_{i=1}^{n} W_i = \frac{1}{250}\sum_{i=1}^{250} W_i = 61977.12.$
  2. $SE(\bar{W}) = \frac{s_W}{\sqrt{n}} = \frac{s_W}{\sqrt{250}} = 1334.19.$
  3. Compute $t^{act} = \frac{\bar{W} - \mu_{W,0}}{SE(\bar{W})} = \frac{61977.12 - 60000}{1334.19} = 1.4819.$
  4. Since we use a 5% significance level, we do not reject H_0 because $|t^{act}| = 1.4819 < z_{0.025} = 1.96.$
- Suppose we are interested in the alternative $H_1 : \mathrm{E}(W) > 60000$. The t-stat is **exactly** the same: $t^{act} = 1.4819$. but now needs to be compared with $z_{0.05} = 1.645$.

 Hypothesis Tests for the Population Mean

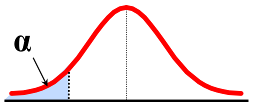
Hypothesis Test for the Mean with σ^2 unknown (n small)

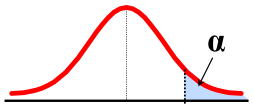
Decision Rules

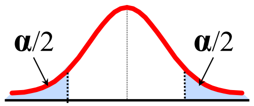
- Consider a random sample of n observations from a population that is normally distributed, ***AND*** variance σ_Y^2 is unknown: $Y_i \sim N(\mu_Y, \sigma_Y^2)$
- Converting the sample average ($\bar{y}$) to a t-value...

Hypothesis Tests for $E(Y)$ $\quad t = \dfrac{\bar{Y} - \mu_{Y,0}}{\mathrm{SE}(\bar{Y})} = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}} \sim t_{n-1}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(Y) \geq \mu_0$ H_1: $E(Y) < \mu_0$	H_0: $E(Y) \leq \mu_0$ H_1: $E(Y) > \mu_0$	H_0: $E(Y) = \mu_0$ H_1: $E(Y) \neq \mu_0$







$-t_{n-1,\alpha}$	$t_{n-1,\alpha}$	$-t_{n-1,\alpha/2}$ $\quad t_{n-1,\alpha/2}$
Reject H_0 if $t < -t_{n-1,\alpha}$	Reject H_0 if $t > t_{n-1,\alpha}$	Reject H_0 if $t < -t_{n-1,\alpha/2}$ or $t > t_{n-1,\alpha/2}$

 Hypothesis Tests for the Population Mean

Hypothesis Test for the Mean with σ^2 unknown (n small)

Example

- The average cost of a hotel room in New York is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{y} = \$172.50$ and $s_y = \$15.40$. Perform a test at the $\alpha = 0.05$ level (assuming the population distribution is normal).
  - Form the hypothesis to be tested

$$H_0 : \mathrm{E}(Y) = 168 \quad \text{the } \textit{mean} \text{ cost } \textbf{is \$168}$$
$$H_1 : \mathrm{E}(Y) \neq 168 \quad \text{the } \textit{mean} \text{ cost } \textbf{is not \$168}$$

  - For $\alpha = 0.05$, with $n = 25$, $t_{n-1,\alpha/2} = t_{24,0.025} = 2.0639$ and $-t_{24,0.025} = 2.0639$, so we would reject H_0 if $|t| > 2.0639$.
  - We have $\bar{y} = 172.50$ and $s_y = 15.40$, so the test statistic is

$$t = \frac{\bar{y} - \mu_{Y,0}}{s_y/\sqrt{n}} = \frac{172.50 - 168}{15.40/\sqrt{25}} = 1.46 < t_{24,0.025} = 2.0639$$

  or $|t| = 1.46 < 2.0639$. Hence H_0 **cannot be** rejected. We **conclude** that there is not sufficient evidence that the true mean cost is different than \$168.

Confidence Intervals for the Population Mean – I

- Suppose we would do a two-sided hypothesis test for many different values of $\mu_{0,Y}$. On the basis of this we can construct a set of values which are not rejected at 5% (α%) significance level.
- If we were able to test all possible values of $\mu_{0,Y}$ we could construct a 95% ($(1-\alpha)$%) confidence interval

Definition

A 95% ($(1-\alpha)$%) confidence interval is an interval that contains the true value of μ_Y in 95% ($(1-\alpha)$%) of all possible random samples.

- A relative frequency interpretation: From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true population mean

Confidence Intervals for the Population Mean – II

- The general formula for all confidence intervals is

$$\text{Point Estimate} \pm \underbrace{(\text{Reliability Factor})(\text{Standard Error})}_{\text{Margin of Error}}$$

$$\hat{\mu} \pm c \cdot \mathrm{SE}(\hat{\mu})$$

and using the sample average estimator

$$\bar{Y} \pm c \cdot \mathrm{SE}(\bar{Y})$$

- Instead of doing infinitely many hypothesis tests we can compute the 95% ($(1-\alpha)$%) confidence interval as

$$\bar{Y} - z_{\alpha/2}\mathrm{SE}(\bar{Y}) < \mu < \bar{Y} + z_{\alpha/2}\mathrm{SE}(\bar{Y}) \quad \text{or} \quad \bar{Y} \pm \underbrace{z_{\alpha/2}\mathrm{SE}(\bar{Y})}_{\text{Margin of Error}}$$

Confidence Intervals for the Population Mean – III

- When the sample size n is large (or when the population is normal and σ_Y^2 is known):
  - A 90% confidence interval for μ_Y: $[\bar{Y} \pm 1.645 \cdot \mathrm{SE}(\bar{Y})]$
  - A 95% confidence interval for μ_Y: $[\bar{Y} \pm 1.96 \cdot \mathrm{SE}(\bar{Y})]$
  - A 99% confidence interval for μ_Y: $[\bar{Y} \pm 2.58 \cdot \mathrm{SE}(\bar{Y})]$
  - with $\mathrm{SE}(\bar{Y}) = \sigma_Y/\sqrt{n}$ when variance is known or $\mathrm{SE}(\bar{Y}) = s_Y/\sqrt{n}$ when unknown and is estimated.

Confidence Intervals for the Population Mean – IV

Example

A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Determine a 95% C.I. for the true mean resistance of the population.

$$\begin{aligned} \bar{y} \pm z_{\alpha/2}\frac{\sigma_Y}{\sqrt{n}} &= 2.20 \pm 1.96(0.35/\sqrt{11}) = 2.20 \pm 0.2068 \\ 1.9932 &< \mu_Y < 2.4068 \end{aligned}$$

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

Confidence Intervals for the Population Mean – V

Example

Using the sample of $n = 250$ individuals with a Ph.D. degree discussed above ($\bar{W} = 61977.12, s_W = 21095.37, \mathrm{SE}(\bar{Y}) = s_W/\sqrt{n} = 21095.37/\sqrt{250}$):

- A 90% C.I. for μ_W is: $[61977.12 \pm 1.64 \cdot 1334.19] = [59349.39, 64604.85]$.
- A 95% C.I. for μ_W is: $[61977.12 \pm 1.96 \cdot 1334.19] = [59774.38, 64179.86]$.
- A 99% C.I. for μ_W is: $[61977.12 \pm 2.58 \cdot 1334.19] = [58513.94, 65440.30]$.

Confidence Intervals for the Population Mean – VI

- When the sample size n is small ***AND*** the population from which we draw data is normal:

$$\bar{Y} - t_{n-1,\alpha/2}\frac{s_Y}{\sqrt{n}} < \mu_Y < \bar{Y} + t_{n-1,\alpha/2}\frac{s_Y}{\sqrt{n}} \quad \text{or} \quad \bar{Y} \pm \underbrace{t_{n-1,\alpha/2}\frac{s_Y}{\sqrt{n}}}_{\text{Margin of Error}}$$

  - A 90% confidence interval for μ_Y: $[\bar{Y} \pm t_{n-1,0.05} \cdot \mathrm{SE}(\bar{Y})]$
  - A 95% confidence interval for μ_Y: $[\bar{Y} \pm t_{n-1,0.025} \cdot \mathrm{SE}(\bar{Y})]$
  - A 99% confidence interval for μ_Y: $[\bar{Y} \pm t_{n-1,0.005} \cdot \mathrm{SE}(\bar{Y})]$
  - with $\mathrm{SE}(\bar{Y}) = s_Y/\sqrt{n}$

Confidence Intervals for the Population Mean – VII

Example

A random sample of $n = 25$ has $\bar{x} = 50$ and $s = 8$. Form a 95% confidence interval for μ.

- $d.f. = n - 1 = 24$, so $t_{24,\alpha/2} = t_{24,0.025} = 2.0639$

$$\begin{aligned} \bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} &= 50 \pm 2.0639(8/\sqrt{25}) = 50 \pm 3.302 \\ 46.698 &< \mu < 53.302 \end{aligned}$$

Comparing Means from Different Populations – I

Large Samples or Known Variances from Normal Populations

- Suppose we would like to test whether the mean wages of men and women with a Ph.D. degree differ by an amount d_0:

$$H_0 : \mu_{W,M} - \mu_{W,F} = d_0 \qquad H_0 : \mu_{W,M} - \mu_{W,F} \neq d_0$$

- To test the null hypothesis against the two-sided alternative we follow the 4 steps as above with some adjustments

1. Estimate $(\mu_{W,M} - \mu_{W,F})$ by $(\bar{W}_M - \bar{W}_M)$.
   - Because a weighted average of 2 independent normal random variables is itself normally distributed we have (using the CLT and the fact that $\mathrm{Cov}(\bar{W}_M, \bar{W}_F) = 0$)

$$\bar{W}_M - \bar{W}_F \sim N\left(\mu_{W,M} - \mu_{W,F}, \frac{\sigma^2_{W,M}}{n_M} + \frac{\sigma^2_{W,F}}{n_F}\right)$$

Comparing Means from Different Populations – II

Large Samples or Known Variances from Normal Populations

2. Estimate $\sigma_{W,M}$ and $\sigma_{W,F}$ to obtain $\text{SE}(\bar{W}_M - \bar{W}_F)$:

$$\text{SE}(\bar{W}_M - \bar{W}_F) = \sqrt{\frac{s^2_{W,M}}{n_M} + \frac{s^2_{W,F}}{n_F}}$$

3. Compute the t-statistic

$$t^{act} = \frac{(\bar{W}_M - \bar{W}_M) - d_0}{\text{SE}(\bar{W}_M - \bar{W}_F)}$$

4. Reject H_0 at a 5% significance level if $|t^{act}| > 1.96$ or if the p-value< 0.05.

Comparing Means from Different Populations – III

Large Samples or Known Variances from Normal Populations

Example

Suppose we have random samples of 500 men and 500 women with a Ph.D. degree and we would like to test that the mean wages are equal:

$$H_0 : \mu_{W,M} - \mu_{W,M} = 0 \quad H_1 : \mu_{W,M} - \mu_{W,M} \neq 0$$

We obtained $\bar{W}_M = 64159.45$, $\bar{W}_F = 53163.41$, $s_{W,M} = 18957.26$, and $s_{W,F} = 20255.89$. We have:

1. $\bar{W}_M - \bar{W}_F = 64159.45 - 53163.41 = 10996.04$.
2. $\text{SE}(\bar{W}_M - \bar{W}_F) = 1240.709$.
3. $t^{act} = \frac{(\bar{W}_M - \bar{W}_F) - 0}{\text{SE}(\bar{W}_M - \bar{W}_F)} = \frac{10996.04}{1240.709} = 8.86$.
4. Since we use a 5% significance level, we reject H_0 because $|t^{act}| = 8.86 > 1.96$

Confidence Interval for the Difference in Population Means

- The method for constructing a confidence interval for 1 population mean can be easily extended to the difference between 2 population means.
- A hypothesized value of the difference in means d_0 will be rejected if $|t| > 1.96$ and will be in the confidence set if $|t| \leq 1.96$.
- Thus the 95% confidence interval for $\mu_{W,M} - \mu_{W,F}$ are the values of d_0 within ± 1.96 standard errors of $(\bar{W}_M - \bar{W}_F)$.
- So a 95% confidence interval for $\mu_{W,M} - \mu_{W,F}$ is

$$\begin{aligned} &(\bar{W}_M - \bar{W}_M) \pm 1.96 \cdot \text{SE}(\bar{W}_M - \bar{W}_M) \\ &10996.04 \pm 1.96 \cdot 1240.709 \\ &[8561.34, 13430.73] \end{aligned}$$

Testing Population Mean Differences

Normal Populations, **Unknown Variances** σ^2_X and σ^2_Y but Assumed **Equal**

$$t = \frac{(\bar{X} - \bar{Y}) - d_0}{\text{SE}(\bar{X} - \bar{Y})} = \frac{(\bar{X} - \bar{Y}) - d_0}{\sqrt{(s^2_p/n_X) + (s^2_p/n_Y)}} \sim t_{n_X+n_Y-2};$$

$$\text{where } s^2_p = \frac{(n_X - 1)s^2_X + (n_Y - 1)s^2_Y}{n_X + n_Y - 2}$$

- The C.I. is constructed as $(\bar{X} - \bar{Y}) \pm t_{n_X+n_Y-2,\alpha/2} \cdot \text{SE}(\bar{X} - \bar{Y})$.
- Recall $\mu_X = \text{E}(X)$, $\mu_Y = \text{E}(Y)$

Lower-tail	Upper-tail	Two-tailed
$H_0 : \mu_X - \mu_Y \geq d_0$ $H_1 : \mu_X - \mu_Y < d_0$	$H_0 : \mu_X - \mu_Y \leq d_0$ $H_1 : \mu_X - \mu_Y > d_0$	$H_0 : \mu_X - \mu_Y = d_0$ $H_1 : \mu_X - \mu_Y \neq d_0$
Reject H_0 if $t < t_\alpha$	Reject H_0 if $t > t_\alpha$	Reject H_0 if $|t| > t_{\alpha/2}$

Testing Population Mean Differences – I

Example: Normal Populations, **Unknown Variances** σ_X^2 and σ_Y^2 but Assumed **Equal**

- You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number:	21	25
Sample mean:	3.27	2.53
Sample std. dev.:	1.30	1.16

Assuming both populations are approximately normal with equal variances, is there a difference in average yield ($\alpha = 0.05$)?

- The hypothesis of interest is

$$\boxed{\begin{array}{l} H_0 : \mu_{NYSE} - \mu_{NASDAQ} = 0 \\ H_1 : \mu_{NYSE} - \mu_{NASDAQ} \neq 0 \end{array}} \text{ or } \boxed{\begin{array}{l} H_0 : \mu_{NYSE} = \mu_{NASDAQ} \\ H_1 : \mu_{NYSE} \neq \mu_{NASDAQ} \end{array}}$$

Testing Population Mean Differences – II

Example: Normal Populations, **Unknown Variances** σ_X^2 and σ_Y^2 but Assumed **Equal**

- Note that $df = n_X + n_Y - 2 = 21 + 25 - 2 = 44$, so the critical value for the test is $t_{44,0.025} = 2.0154$
- The pooled variance is:

$$s_p^2 = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2} = \frac{(21-1)1.30^2 + (25-1)1.16^2}{(21-1)+(25-1)} = 1.5021$$

- The test statistic is

$$t^{act} = \frac{(\bar{x}-\bar{y})-d_0}{\sqrt{(s_p^2/n_X)+(s_p^2/n_Y)}} = \frac{(3.27-2.53)-0}{\sqrt{1.5021\left(\frac{1}{21}+\frac{1}{25}\right)}} = 2.040.$$

Since $|t^{act}| > t_{44,0.025} = 2.0154$, we reject H_0 at $\alpha = 0.05$. We conclude that there is evidence of a difference...

- The C.I. is constructed as $(\bar{X}-\bar{Y}) \pm t_{n_X+n_Y-2,\alpha/2} \cdot \mathrm{SE}(\bar{X}-\bar{Y})$

Testing Population Mean Differences – I

Matched or Paired Samples

- Suppose we obtain a sample of n observations from two populations which are normally distributed and we have paired or matched samples – repeated measures (before/after).
- Define, the pair difference $d_i = X_i - Y_i$. We have

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{X} - \bar{Y}; \quad \text{and} \quad S_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$$

with $\mathrm{E}(\bar{d}) = \mu_d = \mathrm{E}(X) - \mathrm{E}(Y)$ and $\mathrm{SE}(\bar{d}) = \sqrt{\frac{S_d^2}{n}} = S_d/\sqrt{n}$

- If the sample size is large enough ($n \to \infty$) then

$$\frac{\bar{d}-\mu_d}{S_d/\sqrt{n}} \sim N\left(0, \frac{S_d^2}{n}\right).$$

If the sample size is relatively small, then

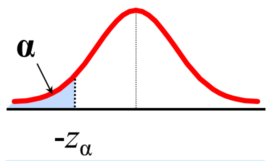
$$\frac{\bar{d}-\mu_d}{S_d/\sqrt{n}} \sim t_{n-1}.$$

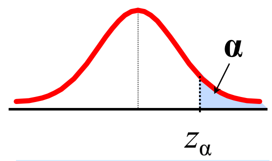
Testing Population Mean Differences – II

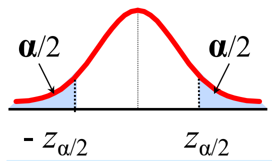
Matched or Paired Samples

Matched or Paired Samples $t = \dfrac{\bar{d}-d_0}{\mathrm{SE}(d)} = \dfrac{\bar{d}-d_0}{s_d/\sqrt{n}}$ (n large)

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(X)–E(Y) \geq 0$ H_1: $E(X)–E(Y) < 0$	H_0: $E(X)–E(Y) \leq 0$ H_1: $E(X)–E(Y) > 0$	H_0: $E(X)–E(Y) = 0$ H_1: $E(X)–E(Y) \neq 0$







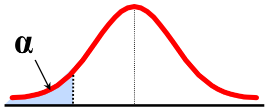
Reject H_0 if $t < -z_\alpha$	Reject H_0 if $t > z_\alpha$	Reject H_0 if $t < -z_{a/2}$ or $t > z_{a/2}$

Testing Population Mean Differences – III

Matched or Paired Samples

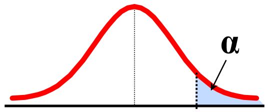
Matched or Paired Samples $t = \frac{\bar{d} - d_0}{\mathrm{SE}(d)} = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} \sim t_{n-1}$

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $E(X)–E(Y) \geq 0$ H_1: $E(X)–E(Y) < 0$	H_0: $E(X)–E(Y) \leq 0$ H_1: $E(X)–E(Y) > 0$	H_0: $E(X)–E(Y) = 0$ H_1: $E(X)–E(Y) \neq 0$



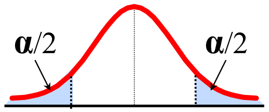
$-t_{n-1,\alpha}$

Reject H_0 if $t < -t_{n-1,\alpha}$



$t_{n-1,\alpha}$

Reject H_0 if $t > t_{n-1,\alpha}$



$-t_{n-1,\alpha/2}$ $t_{n-1,\alpha/2}$

Reject H_0 if $t < -t_{n-1,a/2}$ or $t > t_{n-1,a/2}$

Testing Population Mean Differences – I

Matched or Paired Samples: Example

- Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? Test at the 5% significance level. You collect the following data:

Salesperson	C.B.	T.F	M.H.	R.K.	M.O.
Complaints, Before:	6	20	3	0	4
Complaints, After:	4	6	2	0	0
Difference, d_i	-2	-14	-1	0	-4

$$\bar{d} = \frac{1}{5}\sum_{i=1}^{5} d_i = -4.2; \; s_d = \sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(d_i - \bar{d})^2} = 5.67$$

- The hypothesis of interest is

$$H_0 : \mu_X - \mu_Y = 0$$
$$H_1 : \mu_X - \mu_Y \neq 0$$

Testing Population Mean Differences – II

Matched or Paired Samples: Example

- With $n = 4$ and $\alpha = 0.05$ the critical value is $t_{n-1,\alpha/2} = t_{4,0.025} = 2.776$.
- We have

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{4}} = -1.66 > -t_{4,0.025} = -2.776,$$

or $|t| < t_{4,0.025} = 2.776$. Hence, we **do not reject** H_0. There is not a significant change in the number of complaints.

Annex: Hypothesis Tests – I

Employing the *p*-value

- Suppose we have a sample of n observations (they are assumed *IID*) and compute the sample average $\bar{Y}$. The sample average can differ from $\mu_{Y,0}$ for two reasons
  1. The population mean μ_Y is not equal to $\mu_{Y,0}$ (H_0 is not true)
  2. Due to random sampling $\bar{Y} \neq \mu_Y = \mu_{Y,0}$ (H_0 is true)
- To quantify the second reason we define the *p*-value. The ***p-value*** is the probability of drawing a sample with $\bar{Y}$ at least as far from $\mu_{Y,0}$ as the value actually observed, given that the null hypothesis is true.

$$p\text{-value} = \Pr_{H_0}\left[\left|\bar{Y} - \mu_{Y,0}\right| > \left|\bar{Y}^{act} - \mu_{Y,0}\right|\right],$$

where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed

Annex: Hypothesis Tests – II

Employing the p-value

- To compute the p-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if n is small. With large n the CLT states that

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right),$$

which implies that if the null hypothesis is true:

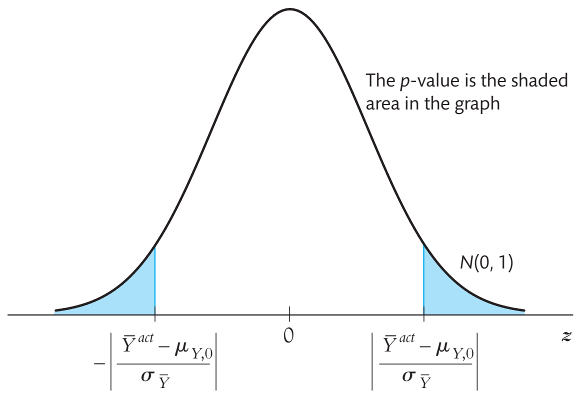
$$\frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}} \sim N(0,1)$$

- Hence

$$p\text{-value} = \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}}\right|\right] = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sqrt{\frac{\sigma_Y^2}{n}}}\right|\right)$$

Annex: Hypothesis Tests – III

Employing the p-value



- For large n, p-value = the probability that a $N(0,1)$ random variable falls outside $\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|$, where $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$

Annex: Hypothesis Tests – I

Computing the p-value when σ_Y^2 is unknown

- In practice σ_Y^2 is usually unknown and must be estimated
- The sample variance S_Y^2 is the estimator of $\sigma_Y^2 = \mathrm{E}\left[(Y - \mu_Y)^2\right]$, defined as

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  - division by $n-1$ because we 'replace' μ_Y by $\bar{Y}$ which uses up 1 degree of freedom
  - if $Y_1, Y_2, \ldots, Y_n$ are *IID* and $\mathrm{E}(Y^4) < \infty$, then $S_Y^2 \xrightarrow{p} \sigma_Y^2$ (Law of Large Numbers)
- The sample standard deviation $S_Y = \sqrt{S_Y^2}$, is the estimator of σ_Y.

Annex: Hypothesis Tests – II

Computing the p-value when σ_Y^2 is unknown

- The standard error $SE(\bar{Y})$ is an estimator of $\sigma_{\bar{Y}}$

$$SE(\bar{Y}) = \frac{S_Y}{\sqrt{n}}$$

- Because S_Y^2 is a consistent estimator of σ_Y^2 we can (for large n) replace

$$\sqrt{\frac{\sigma_Y^2}{n}} \text{ by } SE(\bar{Y}) = \frac{S_Y}{\sqrt{n}}$$

- This implies that when σ_Y^2 is unknown and $Y_1, Y_2, \ldots, Y_n$ are *IID* the p-value is computed as

$$p-\text{value} = 2\Phi\left(-\left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}\right|\right)$$

Statistics for Business

Correlation and Regression

Panagiotis Th. Konstantinou

MSc in International Shipping, Finance and Management,

Athens University of Economics and Business

First Draft: August 20, 2016. **This Draft**: August 28, 2023.

Regression: Examples

- Let y be a student's college achievement, measured by his/her GPA. This might be a function of several variables:
  - x_1 = rank in high school class
  - x_2 = high school's overall rating
  - x_3 = high school GPA
  - x_4 = SAT scores
  - We want to predict y using knowledge of x_1, x_2, x_3 and x_4.
- Let y be the monthly sales revenue for a company. This might be a function of several variables:
  - x_1 = advertising expenditure
  - x_2 = time of year
  - x_3 = state of economy
  - x_4 = size of inventory
  - We want to predict y using knowledge of x_1, x_2, x_3 and x_4.

Regression: A Two Variable Model – I

- If we want to describe the relationship between y and x for the **whole population**, there are two models we can choose
  - Deterministic Model:

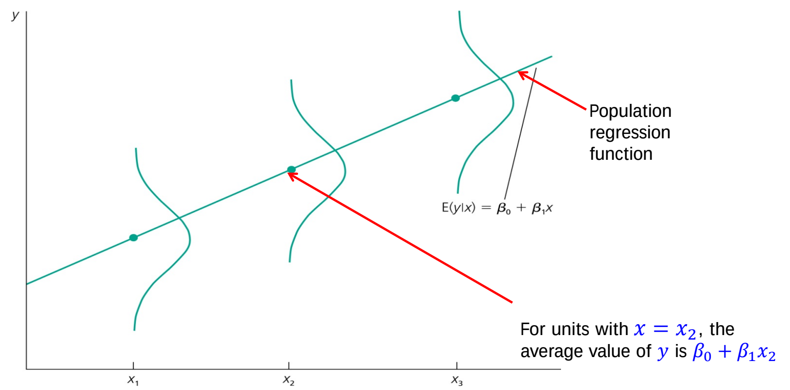
$$\underbrace{y}_{\text{Dependent}} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \underbrace{x}_{\text{Independent}} .$$

  - Probabilistic Model:

$$\begin{aligned} y &= \text{Deterministic Model} + \text{Random Error} \\ y &= \beta_0 + \beta_1 x + \varepsilon. \end{aligned}$$

Regression: A Two Variable Model – II

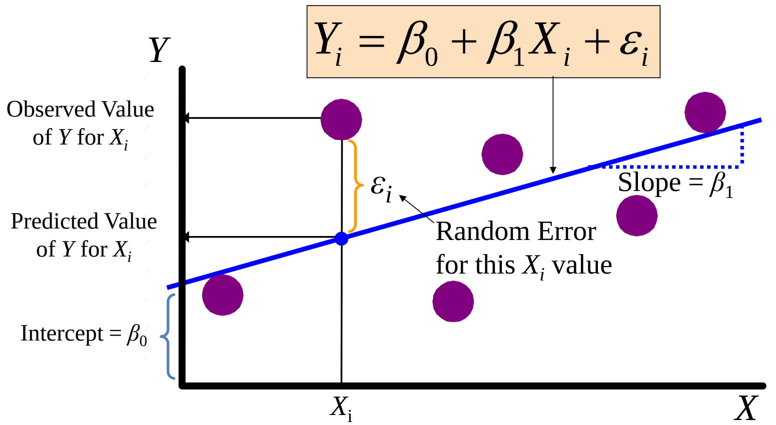
- Since the bivariate measurements that we observe do not generally fall exactly on a straight line, we choose to use a **probabilistic model**.



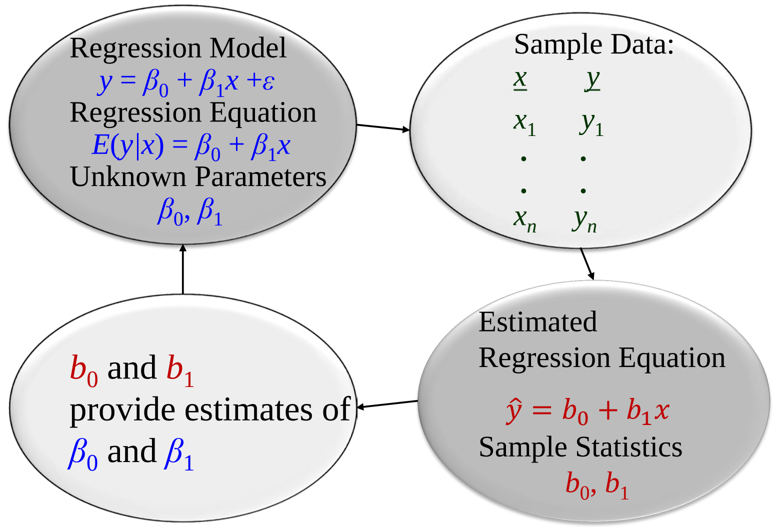
- Points deviate from the population regression line (line of means) by an amount ε, where $\varepsilon \sim N(0, \sigma^2)$.

Regression: A Two Variable Model – III

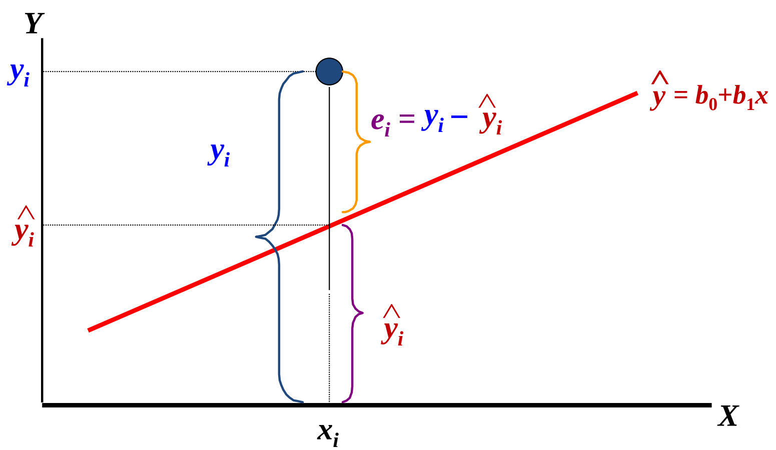
- The population of measurements is generated as y deviates from the population line by ε.



Regression: Estimation Process



Regression Equation and LS – I



Regression Equation and LS – II

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared differences** between y_i and $\hat{y}_i$:

$$\begin{aligned}
\min SSE &= \min \sum_{i=1}^{n} e_i^2 \\
&= \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \min \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2
\end{aligned}$$

Regression Equation and LS – III

- Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\widehat{\mathrm{Cov}(x,y)}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

- The (sample) regression line always goes through the means $\bar{x}$, $\bar{y}$.

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values)
- b_1 is the estimated change in the average value of y as a result of a one-unit change in x :

$$\Delta y = b_1 \Delta x \text{ so}$$

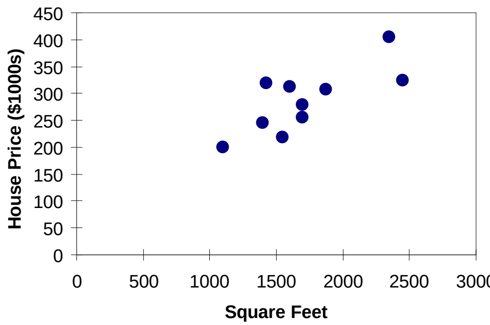
$$b_1 = \frac{\Delta y}{\Delta x}$$

Simple Linear Regression – I

An Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in \$1000s
  - Independent variable (X) = square feet

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Simple Linear Regression – II

An Example

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.762113713					
5	R Square	0.580817312					
6	Adjusted R Square	0.528419476					
7	Standard Error	41.33032365					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	18934.9348	18934.9348	11.0848	0.01039	
13	Residual	8	13665.5652	1708.1957			
14	Total	9	32600.5				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	98.24833	58.03348	1.69296	0.12892	-35.57711	232.07377
18	Square Feet (X)	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression – III

An Example

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

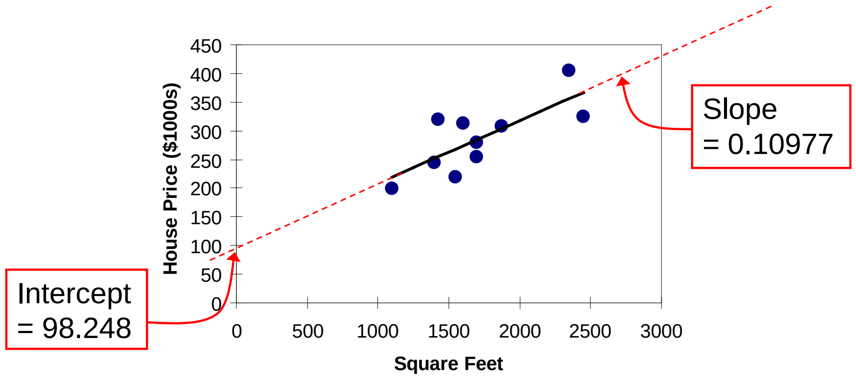
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Simple Linear Regression – IV

An Example



$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Simple Linear Regression – V

An Example

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet}).$$

- b_0 is the estimated average value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
  - Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, $\$98,248.33$ is the portion of the house price not explained by square feet.
- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by $.10977(\$1000) = \$109.77$, on average, for each additional one square foot of size.

Error Variance Estimation – I

- An estimator for the variance of the population model error is

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SSE}{n-2}.$$

  - Division by $n-2$ instead of $n-1$ is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one
  - The **standard error of the estimate** or the **standard error of the regression** is simply

$$SER = s_e = \hat{\sigma} = \sqrt{s_e^2}.$$

Error Variance Estimation – II

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$\boxed{S_e = 41.33032}$$

ANOVA

	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Prediction – I

- Recall from our discussion above that the **fitted** or **predicted** value for observation i is

$$Y_i = b_0 + b_1 X_i.$$

- Given that we have estimated the parameters of the model (and assessed its statistical significance) we may want to:
  - Estimate the average value of Y at a given value of $X = X_0$;
  - Predict a particular value of Y for a given value of $X = X_0$.
- In both cases the point estimate is

$$\hat{Y}_0 = b_0 + b_1 X_0.$$

Prediction – II

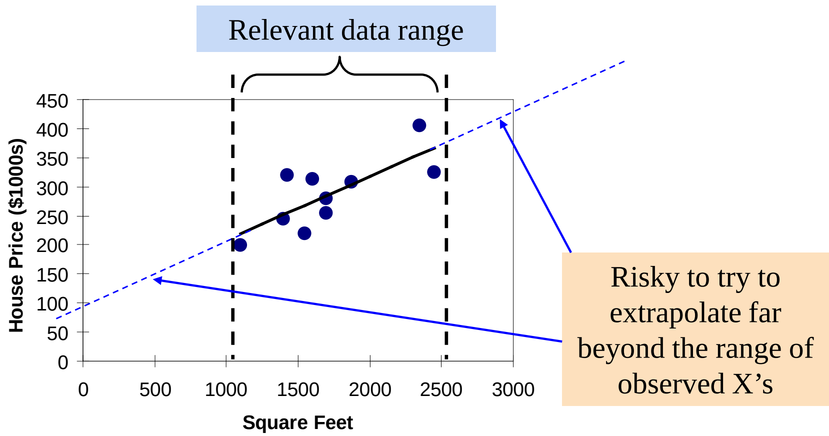
- Predict the price for a house with 2000 square feet:

$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098 \cdot (\text{square feet}) \\ &= 98.25 + 0.1098 \cdot (2000) \\ &= 317.85\end{aligned}$$

- The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850.

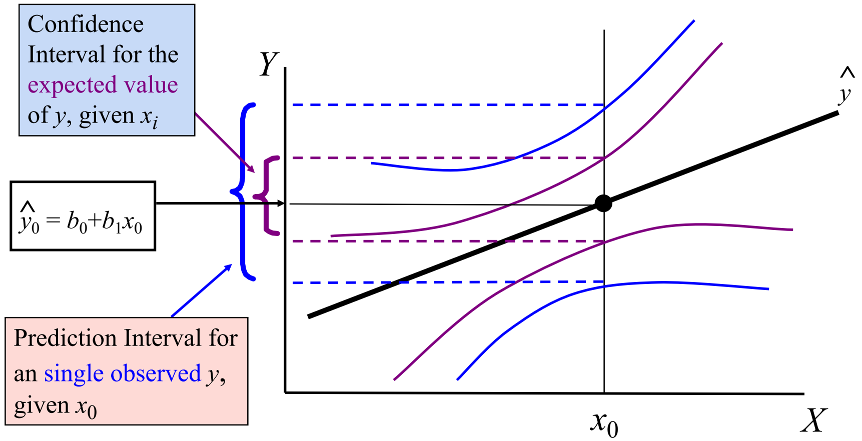
Prediction – III

- When using a regression model for prediction, only predict within the relevant range of data



Prediction – IV

- *Goal*: Form intervals around Y to express uncertainty about the value of Y_0 for a given X_0



Prediction – V

- Confidence interval estimate for the expected value of y given a particular x_0

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  - Notice that the formula involves the term $(x_0 - \bar{x})^2$ so the size of interval varies according to the distance x_0 is from the mean, $\bar{x}$.
  - Technically this formula is used for infinitely large populations. However, we can interpret our problem as attempting to determine the average selling price of **all** houses, all with 1,500 square feet.

Prediction – VI

- Confidence interval estimate for an actual observed value of y given a particular x_0

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  - The extra term (1) comes in because the regression is used to estimate the value of **one value** of y (at given x_0)
- Confidence Interval Estimate for $E(Y_0|X_0)$: Find the 95% confidence interval for the mean price of 2,000 square-foot houses
  - Predicted Price $\hat{y} = 317.85(\$1,000s)$ so

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 317.84 \pm 37.15$$

Prediction – VII

  - The confidence interval endpoints are 280.66 and 354.90, or from \$280,660 to \$354,900
- Confidence Interval Estimate for $\hat{Y}_0$: Find the 95% confidence interval for an individual house with 2,000 square feet
  - Predicted Price $\hat{y} = 317.85(\$1,000s)$ so

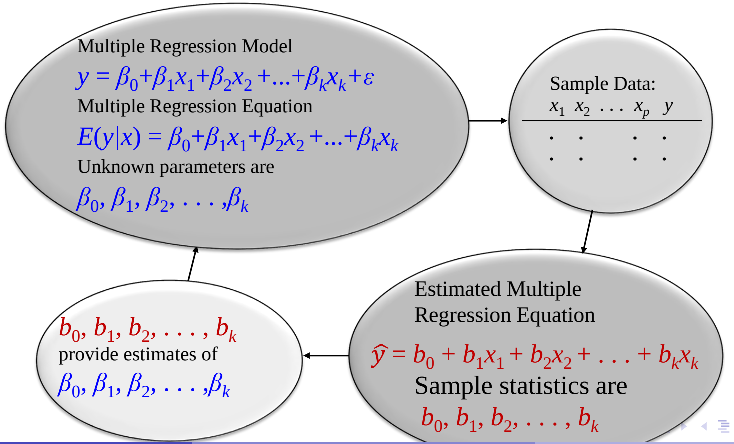
$$\hat{y}_0 \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 317.84 \pm 102.28$$

  - The confidence interval endpoints are 215.50 and 420.07, or from \$215,500 to \$420,070.

Multiple Regression

- If we want to describe the relationship between one dependent variable y and two or more independent ones $x_1, x_2, \ldots, x_k$ for the **whole population**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$



Multiple Regression: An Example – I

- A distributor of frozen desert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in$) Advertising ($100's)
  - Data are collected for 15 weeks

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

- Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1(\text{Price}) + b_2(\text{Advertising})$$

Multiple Regression: An Example – II

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Regression: An Example – III

- The estimated multiple regression equation

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

  - $b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising (assuming these do not change)
  - $b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price (assuming these do not change).

Multiple Regression: Prediction – I

- Let a population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i;$$

then given a new observation of a data point

$$x_{1,n+1}, x_{2,n+1}, \cdots, x_{k,n+1}$$

the best linear, unbiased forecast of y_{n+1} is

$$\hat{y}_i = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_k x_{k,n+1}$$

  - It is risky to forecast for new x values outside the range of the data used to estimate the model coefficients, because we do not have data to support that the linear model extends beyond the observed range.

Multiple Regression: Prediction – II

- Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\begin{aligned} \widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62 \end{aligned}$$

  - Note that Advertising is in \$100's, so \$350 means that $x_2 = 3.5$.
  - Predicted sales is 428.62 pies